2023

A Machine learning tool to predict early-stage start-up success in Africa.

Gichohi, Brian Waihiga
School of Computing and Engineering Sciences
Strathmore University

A Machine Learning Tool to Predict Early-Stage Start-Up Success in Africa

Brian Waihiga Gichohi

138134

Master of Science in Computing and Information Systems

2023

A Machine Learning Tool to Predict Early-Stage Start-Up Success in Africa

Brian Waihiga Gichohi

138134

Submitted in partial fulfillment of the requirements for the Degree of

Master of Science in Computing and Information Systems at Strathmore University

School of Computing and Engineering Sciences

Strathmore University

Nairobi, Kenya

June 2023

Declaration

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

Student's Name: Brian Waihiga Gichohi

Sign: _____BWG_____ Date: 22nd May 2023

Approval

This dissertation was done by Brian Waihiga Gichohi and was reviewed and approved by the following:

Dr. Bernard Shibwabo
Senior Lecturer,
School of Computing and Engineering Sciences,
Strathmore University

Dr. Julius Butime,
Dean, School of Computing & Engineering Sciences,
Strathmore University

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

Abstract

Most start-ups do not celebrate their first year in operation, and a few survive to see their fifth year of operation. This has been a challenge for all the stakeholders involved. Therefore, an effective tool for predicting the possibility of a start-up surviving its infancy stages and eventually growing into a profitable venture could be a breakthrough for entrepreneurs, innovators, and investors. This study assessed the factors that make early-stage start-ups successful, specifically in Africa and developed a web-based prototype that uses machine learning algorithms to predict the success of proposed start-ups. The study adopted both descriptive research design and applied research. Data was collected using a secondary data source called CrunchBase, a global investor platform. This data formed the basis for the development of the prediction tool. The tool was designed to predict the success or failure of start-ups based on the collected data. To ensure the accuracy and reliability of the prediction model, 80% of the collected data was used for training the model, while the remaining 20% was utilized for testing and validation purposes. The model development employed Artificial Neural Networks (ANNs) algorithm, known for its capability to analyze complex patterns and relationships in data. The developed model achieved an impressive accuracy of 86.81%, indicating its effectiveness in predicting the success of start-ups. The tool was implemented using Flask, a Python web framework, along with other Python machine learning frameworks such as Keras and Sci-kit Learn. This allowed for the development of a user-friendly and interactive web-based prototype. A number of users were provided access to the tool for usability testing, and their feedback indicated that the tool was intuitive, easy to use, and effective in predicting the success of start-ups. This study successfully developed a web-based prototype using agile methodology, integrating machine learning algorithms based on Artificial Neural Networks. The prototype demonstrated high accuracy in predicting start-up success, making it a valuable tool for entrepreneurs, innovators, and investors in Africa and beyond.

Keywords: Business start-ups, machine learning algorithms, prediction tool, start-up success.

Table of Contents

List of Figures

List of Tables

Abbreviations/ Acronyms

ANN	Artificial Neural Networks
API	Application Programming Interface
AUC	Area Under Curve
CART	Classification and Regression Trees
CRM	Consumer Relationship Management
EU	European Union
FN	False Negative
FP	False Positive
IPO	Initial Public Offering
KIM	Kenya Institute of Management
MLP	Multilayer Perceptron
ROC	Receiver Operating Characteristic
R&D	Research and Development
SMEs	Small and Medium Enterprises
SPSS	Statistical Package for the Social Science
SVM	Support Vector Machine
TP	True positive
TN	True Negative
US	United States
VC	Venture Capitalists

Acknowledgements

First and foremost, I thank the Almighty God for giving me the strength and opportunity to pursue this master's Degree. I also wish to acknowledge and wholeheartedly thank my dissertation supervisor, Dr. Bernard Shibwabo, for his guidance throughout this process. I could not have done it successfully without his help.

Dedication

This dissertation is dedicated to my late father. He was not able to see me complete this master's journey but I know he is very proud. To my lovely wife, Mumbi, and our son Thagicu, thank you for your patience, sacrifice, and support throughout this journey.

Definition of Terms

Application Program Interface	Software intermediary that allows two applications to talk to each other(Park, 2017)
Initial Public Offering	The term "initial public offering" (IPO) refers to the first time a company's stock is sold to the general public. (Danilov, 2016).
Machine Learning	Machine learning refers to a computer's growing capacity to handle data without new instructions (Ratner, 2017).
Small and Medium Enterprise	The government of Kenya (GOK, 2005) defines a microenterprise as a business with between one and ninety-nine employees that operates formally, informally, seasonally, or year-round in various locations, including markets, streets, households, and mobile units.
Start-up	A startup is a temporary enterprise to discover a sustainable business strategy (Blank & Dorf, 2020).
Venture Capitalist	A venture capitalist (VC) is a sort of private equity investor who, in exchange for a stake in the company's future success, provides funds to emerging businesses. (Rogers, 2020).

Chapter 1: Introduction

1.1 Background of the Study

A start-up is an entrepreneurial endeavour to disrupt industries and change the world (Baldridge, 2022). Due to their tremendous growth potential, startups typically have a very high risk and reward profile (Achleitner, 2010). This high growth potential gives rise to scalability, which is crucial for new businesses.

The aspect of business start-up failure is evident. Statistics show that 90% of start-ups fail. Less than 50% of the start-ups survive to see their 5th year, while only 33% survive past their 10th year (Chernev, 2022). Start-up failure in Africa has followed a similar trend of closing shops in their early years of operation. In 2020, the average startup failure rate in Africa was 54%. However, the rate varied between nations. In the same year, 75% of start-ups in Ethiopia and Rwanda ceased operations, while the failure rate for start-ups in Kenya was 24% (Saleh, 2022). Research conducted by the Kenya National Bureau of Statistics revealed that about 400,000 start-ups in Kenya do not survive celebrating their second year of operation, with a few reaching their fifth year. About 46% of the start-ups in Kenya are estimated to collapse within their first year of launching (Wakiaga, 2020).

Both researchers and entrepreneurs have struggled over the years to predict the success of a start-up. Yet, the success of a venture is what founders and investors look forward to when investing in a business venture. As a result, there has been an active search for methods, tools, and advice to determine the possibility of the success of a start-up venture (Żbikowski & Antosiuk, 2021). An effective tool for predicting the likelihood of a start-up surviving the infancy stages and eventually growing into a profitable venture could be a breakthrough for entrepreneurs, innovators, and investors.

The elements needed for a new firm to prosper have been determined via studies. Watson and Hogarth-Scott (1998) conducted an empirical study on start-ups, their survival, and their growth or failures to determine the success criteria. The research concluded that the founder's experience, difficulties managing the start-up, reasons for starting and maintaining the business, and focus on expansion were all critical success factors for start-ups. Okrah, Nepp, and Agbozo (2018) examined the growth and success variables for start-ups. The findings showed that internal market openness, turnover, governmental regulations, and market dynamics affect start-up performance. Entrepreneurs and investors, however, continue to face uncertainty over the success or failure of their proposed endeavours. To

have sustainable entrepreneurial ventures, it is essential to evaluate the risks involved and comprehend the uncertainties surrounding each potential company opportunity (Tomy & Pardede, 2018).

Most startups fail in their early stages because their founders are unable to cope with the risks and implications of the many unknowns that arise at every turn. How founders handle risks before committing to a new endeavour can determine whether or not the business will succeed. (Butler, Doktor, & Lins, 2010). Venture capitalists are also uncertain about which venture to invest their money in. When venture capitalists invest in start-ups, they expect long-term growth, eventually giving them considerable investment returns. However, start-ups establish a product or service in an uncertain environment (Ries, 2011).

This study successfully developed a web-based prototype using agile methodology, integrating machine learning algorithms based on Artificial Neural Networks. The tool was designed to predict the success or failure of start-ups based on the collected data. To ensure the accuracy and reliability of the prediction model, 80% of the collected data was used for training the model, while the remaining 20% was utilized for testing and validation purposes. The model development employed Artificial Neural Networks (ANNs) algorithm, known for its capability to analyze complex patterns and relationships in data.

The developed model achieved an impressive accuracy of 86.81%, indicating its effectiveness in predicting the success of start-ups. The tool was implemented using Flask, a Python web framework, along with other Python machine learning frameworks such as Keras and Sci-kit Learn. This allowed for the development of a user-friendly and interactive web-based prototype. A number of users were provided access to the tool for usability testing, and their feedback indicated that the tool was intuitive, easy to use, and effective in predicting the success of start-ups.

1.2 Problem Statement

Entrepreneurship and Innovations in Africa are rising, with new ideas and businesses coming up every day. Some have eventually become very successful ventures, attracting many investors to the African market. In 2021, over $2 billion was invested in tech start-ups in Africa (Jackson, 2022).

Even with this rising potential, it is worth noting that 75% of the start-ups that have already received investments from venture capitalists still fail, and many of those that survive,

sustain a marginal existence (Picken, 2017). This statistic shows the high level of risk involved in investing in start-ups. Currently, most investors determine the potential success of a start-up by analysing historical financial or operational data. Since most early-stage start-ups cannot provide such information, the available data is qualitative. Entrepreneurs and investors find it challenging to make informed and objective decisions in this context because humans are selective in the information they use and are subject to bias when making decisions.

As a result, machine learning models have been developed to mitigate these risks (Unal & Ceasu, 2019). However, the datasets and factors used have not been African-centred. Thus, the accuracies have been imprecise. Additionally, there lacks a tool to be used by non-technical entrepreneurs and investors to predict the success of early-stage start-up businesses. Therefore, there is a need for more effective methods of selecting potential businesses to invest in.

1.3 Research Objective

The main aim of this study is to develop a machine learning tool to predict the success of early-stage business start-ups in Africa.

1.3.1 Specific Objectives

a) To investigate the challenges hindering the success of early-stage start-up businesses in Africa.
b) To analyze the methods used for predicting the success of business start-ups.
c) To review the existing prediction algorithms and models used in the prediction of business start-ups.
d) To develop a machine learning tool to predict the success of early-stage business start-ups in Africa.
e) To test and validate the tool.

1.4 Research Questions

a) What are the factors influencing the success of an early-stage business start-up in Africa?
b) How is the success of an early-stage business start-up predicted?
c) What prediction algorithms and models are currently being used to predict the success of early-stage business start-ups?

d) How can an early-stage business start-up success prediction tool be developed to predict the success of early-stage business start-ups in Africa?

e) How can the functionality of the tool be tested?

1.5 Justification

Investing in a start-up without the certainty of success poses significant risks to investors and entrepreneurs. Many global investors have erroneously invested their money in failed start-ups, causing huge losses. Despite creative ideas and innovation in Africa, many entrepreneurs have been unable to acquire investors to launch or expand their start-ups to the next level. This is attributed to a lack of certainty concerning whether a start-up can survive its early stage to become a profitable venture offering attractive investment returns. This poses a significant threat to the development of business initiatives. Therefore, establishing a predictive tool that could be used to determine the possibility of success of a start-up at the early stage of development would be a breakthrough for the stakeholders involved.

As part of any investor's due diligence, accurately predicting which start-ups have the best chance of success will mitigate the potential for huge losses. This reduced risk is also beneficial to the start-up community as more willing angel investors will be confident enough to enter the market and invest in start-ups that would have otherwise not gotten a chance to grow.

1.6 Scope and Limitations

The research is limited to identifying the factors determining early-stage start-ups' success in Africa. Therefore, start-up data collected only from Africa trains the machine learning tool. Besides, the study develops a tool to predict the success of early-stage start-ups, but not at any other stage of their growth.

Chapter 2: Literature Review

2.1 Introduction

This chapter examines the relevant studies on predicting startup success. Predicting the success of early-stage ventures is extremely difficult and uncertain because, in many cases, there are only vague concepts, no prototypes, and thus proof of concept is still unresolved. (Dellermann et al., 2017). Moreover, such concepts may not yet have a market, but they have significant growth potential in the future. The chapter also examines the current techniques and methodologies used to predict the early years of a venture's success. Focus will be placed on machine learning and classification techniques in order to comprehend how this issue has been addressed. In addition, the numerous models and frameworks associated with this study will be presented. In addition, the chapter provides a summary of the systems and applications for predicting startup success. On the basis of the assessment, a conceptual model demonstrating how the tool will function is presented.

2.2 Empirical Framework

2.2.1 Challenges Hindering the Success of Early-stage Start-up Businesses in Africa

Access to Finance: Obtaining capital is one of the most significant obstacles for African businesses. Numerous companies are underfunded and lack the resources necessary to cover production costs. According to studies, inadequate capital accounts for 80% of failing enterprises (Amne, 2021). Moreover, many entrepreneurs have little financial expertise in managing their firms and acquiring capital through investments.

Infrastructure is one of the most significant obstacles for entrepreneurs in Africa. Numerous nations struggle with inadequate energy, electricity, roads, and transit networks. This forces small enterprises to increase their production costs, leading to an increase in the price of their goods and services to break even, yet they frequently incur a loss. In addition, many nations are losing investors due to the high expense of operating a business in a location with a dysfunctional infrastructure (Amne, 2021).

Start-up brands typically confront the problem of competing with more established brands regarding service quality and financial resources (Njuguna, 2019). This level of competition might be detrimental to enterprises that are not yet well-established enough to capture a substantial market share. It is natural for consumers to gravitate toward brands that have been in business for an extended period since these brands have earned high customer trust. For new brands to thrive, they must provide consistent and superior service quality. Over

time, their rising consumer base will develop a higher level of trust, and their brand's reputation will increase.

Market Access can be difficult for African start-ups. A significant youth population on the continent has access to the internet and can quickly obtain information. This difficulty demonstrates that insufficient market research can lead to ineffective marketing strategy and execution, which may ultimately result in marketing to the incorrect audience (Amne, 2021). Before launching a new product, businesses must conduct extensive research. This will ensure that they are promoting to the appropriate market audience

2.2.2 Factors Influencing the Success of Early-Stage Business Start-Ups

Various empirical studies have examined the success factors that make start-ups and small businesses succeed. Makarenko et al (2019) studied the success factors of small businesses in Russia. They used economic analysis methods to identify the critical success factors for small enterprises. The results identified the following critical success factor; entrepreneurs have in-depth knowledge of their business, clearly defined goals, innovative ideas; a transparent business system; competent employees; and business flexibility.

Okrah, Nepp, and Agbozo (2018) explored start-up success factors in the dynamic world. The findings revealed that the success of start-ups is affected by internal market openness, turnover, government policies, innovation (R&D), financing, and market dynamics. Skawi´nska and Zalewsk (2020) studied the critical success of 13 countries in the European Union. The authors used a multivariate statistical analysis method to examine the main success factors for European Union (EU) start-ups. The findings revealed five main success factors: development potential, business experience; focus on the market situation; quality and outcomes of institutions and business relations; and access to human capital.

Douglas, Douglas, and Muturi (2017) conducted an explorative study of the critical success factors for SMEs in Kenya. They first ran an in-depth literature review to identify the various identified success factors and developed a questionnaire. The questionnaire was used to collect data from sampled SME owners or senior managers in SMEs in Kenya. They were selected from the membership database for the Kenya Institute of Management (KIM). The sampled respondents expressed their views on the critical factors determining their success. The key success factors identified in the study included consumer relationship management (CRM), good marketing skills, and government policies.

Murimi (2014) examined the factors that affected the success of start-ups operated by youths in Nairobi, Kenya. A descriptive and exploratory research design was employed in the study. The author used purposive sampling to establish a representative sample, and data were collected using a structured interview guide. The results revealed the critical success factors for start-ups as entrepreneurial skills, innovativeness, and proactiveness; government legislation and policies; and availability of funds.

2.2.3 Prediction of Business Success

Predicting the success or failure of businesses has been a topic of academics and researchers for decades due to the impact on many involved parties. However, due to reasonably restrictive assumptions, standard statistical approaches were not adequate for these predictions, and hence new methods, such as machine learning, became extensively researched in the late 1990s (Daubie & Meskens, 2012).

A machine learning model to predict business success was built by Żbikowski and Antosiuk (2021) using three algorithms: Logistic Regression, Support Vector Machines (SVM), and XGBoost. The researchers obtained data of 213,171 companies from Crunchbase and based their success indicators by analyzing where the organization had been acquired, attained Initial Public Offering (IPO) or was in operation, and received Series B funding. The researchers attained an accuracy level of 86% for Logistic Regression, 84% for SVM, and 86% for XGBoost.

Aktan (2011) studied the effectiveness of machine learning algorithms in predicting business failure. The research was conducted on 180 production industry businesses, and the financial ratios were obtained from the annual reports. The author assessed five machine learning techniques: Decision Trees, Bayesian models, Support Vector Machines, Artificial Neural Networks, and K-Nearest Neighbour. The results established that the Decision Tree method was the most efficient method for predicting business failure.

Krishna, Agrawal, and Choudhar (2016) conducted a study to establish a predictive model for start-ups on the critical events in different lifecycles of a start-up business. The researchers proposed a method for predicting a startup's outcome based on several critical factors, including seed funding amount, seed funding duration, Series A funding, and factors contributing to the company's success and failure at each milestone.

The researchers applied a range of data mining classification algorithms, data mining optimizations, and data mining validations to the preprocessed data. Analysis was provided using, among other methods, ADTrees, Random Forest, and Bayesian Networks. Metrics including the area under the precision, ROC curve, and recall were used to gauge the accuracy of their models. They showed how a start-up may use its models to prioritise success-critical criteria.

The developed system lacked in some capacities. The accuracy levels were low and varied across the different development algorithms, attaining a median of 87.6. In addition, the model could not be used by non-technical people, who comprise the most significant percentage of the target group.

Li (2010) conducted research to develop a model for classifying start-ups and to analyze the critical features for start-up company success. The researcher utilized Random Forest and Support Vector Machine to investigate and explain several of the crucial factors determining start-up businesses' success. In addition, they contrasted the effectiveness of various machine learning techniques in predicting investors' startup success (Li, 2020).

The researcher obtained data from Kaggle. The data had information on 22,000 start-ups founded from 1997 to 2014 via CrunchBase. The algorithms used in this paper had better accuracy and nearly perfect precision scores averaging 88.5% and 98.5%, respectively, for Random Forest and SVM algorithms. There is still room for improvement in the accuracy of the models using better algorithms such as deep learning models. This research did not develop a tool for typical entrepreneurs without technical knowledge.

Huang (2016) conducted a study to establish a predictive model to predict a start-up's success using machine learning and network analysis. The author used Random Forest, Adaptive boosting, and Kernel Support Vector Machines. The results revealed that Random Forest predicted start-ups' success with great accuracy.

Table 2.1 compares the other various methods used to predict the success of start-ups.

Table 2.1: Comparison of start-up success prediction tools

Author	Methods	Limitations	Accuracy
Greenberg et al. (2013)	Random Forest Logistic Regression	Feature selection is required for effective analysis	Random Forest=0.6753 Logistic Regression=0.6509
Xie et al., 2014)	Sentiment Analysis	Improper analysis of heterogeneous data	0.81
Sharchilev et al. (2018)	Gradient Boosting Logistic Regression Neural Network	Feature selection required to enhance performance	0.62
Kofanov and Zozul`ov (2018)	Bayesian Network	This method does not use feature selection such as a bag of words and TF-IDF to enhance performance	Not provided
Antretter et al. (2019)	Random Forest and Gradient boosting	It lacks legitimate data analysis	0.76
Gyimah et al. (2019)	Lussier Model	High rate of missing data which impacts performance	0.86
Saura et al. (2019)	Support Vector Machine	Less sample size used for prediction	Not provided
Dellermann et al. (2021)	Machine Learning	The method consumes a lot of time and is costly when a large data set is used	Not provided

Dellermann, Lipusch, Ebel, and Michael (2017) predicted the success of early-stage start-ups using machine and collective intelligence. Among the machine learning techniques employed were Artificial Neural Networks, Support Vector Machines, Logistic Regression, Naive Bayes, and Random Forest. The results demonstrated that a hybrid strategy outperforms machine or human-only prediction.

The study by Unal and Ceasu (2019) proposes a suite of replicable models for forecasting the success of startups using machine learning. Crunchbase supplied the information that was used for their study. They ran the data through it first to check for imbalances and biases in the sampling procedure. They utilized six models to try and foretell how well new businesses will do. Ensemble methods, random forest, and extreme gradient boosting were chosen as the best models based on their 94.1 per cent and 94.5 per cent test set prediction accuracy and 92.2 per cent and 92.91 per cent Area Under Curve (AUC), respectively. First funding lag and firm age are identified as the two most important factors in these models (Unal & Ceasu, 2019).

This paper comes up with a near-perfect model for predicting start-up success. Still, it fails to develop a tool that entrepreneurs or innovators can use to perform the predictions efficiently.

2.3 Theoretical Framework

2.3.1 Definition and Evaluation of Start-ups

The term startup was first used to describe newly formed venture-backed technology companies during the dot-com boom of the late 1990s and early 2000s (Kidder, 2013). A startup is a company that aims to create a business model that can be replicated with ease. (Blank, 2010). According to Blank, a business model's primary characteristic is its ability to capture value.

Startups try to reach key financial benchmarks like break-even and profitability with as little initial investment as possible so they may focus on capitalizing on the substantial growth potential. The transition from a new company to an established one is not always easy to identify. There is no hard and fast rule on when a startup stops being a startup. However, it is typically connected to events like becoming profitable, going public, or being acquired. (Kidder 2013).

Coming public or being acquired requires a business evaluation. Estimating the value of a startup can be challenging due to its typical dearth of historical data, negative cash flows, new products, and a high degree of uncertainty (Grummer, 2013). Consequently, assumptions must be made, and even minor modifications to the conditions can significantly impact the evaluation's outcome.

Venture capitalists and angel investors evaluate start-ups based on intuition, investing preferences, experiences, facts, and required estimations such as industry, business stage, location, and capital (Shepherd, 1999). Table 2.2 summarises several researchers' findings on venture capitalists' due diligence when evaluating early-stage start-ups before investing in them. The main questions that they sought to answer were who ("team"), where ("market"), and what ("product").

Table 2.2: VC start-up evaluation criteria

Authors	Factors
MacMillan et al. (1985)	Market growth Founder's track record and demonstration of leadership Articulation of venture Can the venture be liquidated easily? Size of potential return within 5-10 years The capability of sustaining intense effort Evaluation and reaction to risk Proprietary protection
Shepherd (1999)	Timing Founder's Educational capability Founder's Industry related competence Competitive rivalry Lead time
Zacharakis and Meyer (2000)	Management track record Liquidity of venture Product differentiation Level of innovation Business opportunity and industry potential Product proprietary ownership Alignment of goals between investors and founders

2.3.2 Start-Up Financing

Start-up financing refers to the various options available to a startup to acquire the capital it needs to survive and develop into a well-established and mature business (Kollmann, 2009). Start-up financing has five stages, from the pre-seed stage to the Initial Public Offering stage (IPO). This study directs its attention to the early stage, often called the start-up stageIn this phase, capital is required for the final stage of product development, marketing, and

administrative expenses, as well as to support the business's cash flow needs (Goldberg 2012).

Most start-ups are not yet profitable at this stage (Goldberg 2012). The company must expand to keep up with and, more importantly, facilitate growth. This expansion is usually impossible solely through incoming revenues (Kollmann 2003). As a result, new funding rounds are required, where venture capitalists (VC) come into play.

2.3.3 Measures of Success

There are five factors that can be used to categorize success (Agha, 2014). The first category is the ability of a start-up to reach profitability. Profitability is based on the start-up consistently generating positive cash flow. Along with profitability, the following identified success criterion is self-sufficiency. This means that, aside from the revenues generated by the start-up, no external funds are required to survive in the market. Ideally, the start-up has regular customers using their product or services for new customers and recurring revenue. The impact of a start-up on its surroundings comes next. When a company is profitable and doing good, it is considered successful.

Following that, the main criterion introduced by investors is the exit. A significant increase in sales could generate the effect of an exit for the investors. This means the start-up's sales are so high that it no longer needs to be purchased. The profit is sufficient to pay the investors' returns while keeping the company running (Agha, 2014). Finally, an IPO, or going public on the stock exchange, has been identified as an optional success criterion.

When all these considerations are made, a successful start-up will generate enough revenue to support itself without seeking outside investment, will have a beneficial influence on the lives of at least some people, and will significantly impact the market. Having an exit strategy or going public are also signs of success.

2.4 Current Techniques and Approaches Used for Prediction

2.4.1 Machine Learning

The term "machine learning" was coined by an American computer scientist called A.L Samuel in the late 1950s when he established a self-learning checkers software. However, widespread acceptance of machine learning did not occur until after the 2000s, when advances in computing power and the increased availability of big data accelerated its application. It is a subset of artificial intelligence (Mitchell, 1997).

According to Ratner (2017), Machine learning is a computer's (a machine's) ability to learn data structure without being explicitly programmed. Therefore, algorithms and statistical models are utilized in machine learning to execute different mathematical operations on sampled data and learn rules and relations within them. The main task of machine learning is making inferences from the sampled data, also known as training data. Therefore, machine learning employs statistical theory in developing models representing patterns in data, which, if discovered, can generalize complex issues.

The model can be used for prediction, description, or both (Alpaydin, 2014). While the predictive models predict what will happen in future scenarios based on data that describes what has happened in the past, descriptive models help illustrate the structure of facts and aid in explaining and comprehending a situation (Witten, Frank, Hall, & Pal, 2017). Two main types of machine learning are differentiated due to the different data entry structures employed. These include unsupervised learning and supervised learning.

2.4.1.1 Supervised Learning

Supervised learning necessitates training material consisting of examples of input variables (X) and corresponding output variables (Y) to learn the mapping from the input to the output ($Y=f(X)$). After learning the rule for mapping input to output from the training dataset, the algorithm can predict or classify output from new input data, producing valid predictions for unique cases (Alpaydin, 2014). In supervised learning, the possible outcome from the input is known. The data has all been labelled, and the algorithms are learning to anticipate the output based on the input data. While a regression function generates models where the output variable is a fundamental value such as 80, 150, or 0%, a classification function builds models with a category; for instance, not acquired/acquired, blue/red. The methods used are classification (output is a category) or statistical regression (output is a number).

2.4.1.2 Unsupervised Learning

Unsupervised learning entails a dataset containing input variables without expected or desired output. Unsupervised learning aims to identify the patterns in the input space that happen more than others. The program will identify and classify possible outcomes during the learning phase. In unsupervised learning, the possible answers are unknown since all data is not labelled; hence, the algorithm learns to establish patterns from the inputted data. Unsupervised learning is commonly classified as association and clustering analysis. While the clustering challenge is the discovery of groups with various features between them and

homogenous features between each group's observations, an association problem occurs when one wants to discover the rules that describe a large amount of data; for instance, individuals who acquire A also tend to purchase B (Aggarwal, 2015).

2.4.2 Classification Techniques

In machine learning, classification or regression techniques are among the supervised learning techniques employed on labelled datasets to establish a predictive model for an outcome variable based on various independent input variables (Krishna, Agrawal, & Choudhary, 2016). Some of these techniques are discussed in the following sub-section.

2.4.2.1 Random Forest

The random forest algorithm is widely regarded as one of the most successful machine learning methods currently available, as evidenced by past studies on start-up success prediction. It is an exceptionally versatile tree-based approach, easily adaptable to different settings where other algorithms may not be applicable. Leo Bierman introduced this technique in 2001. The basic principle of random forest is best described as "divide and conquer," where the algorithm continually divides the dataset into fractions, grows a randomized tree predictor on each little piece, and then aggregates these predictors together (Biau & Scornet, 2015). According to Verikas et al. (2016), a random forest is a committee of decision trees where the final output is based on the majority voting. This concept of random forest is illustrated in Figure 2.1.

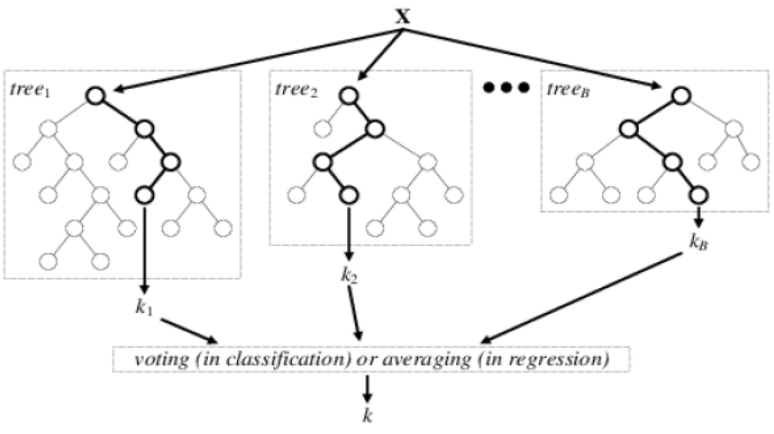


Figure 2.1: Random Forest (Verikas et al., 2016)

Bagging and bootstrap are two essential terms that need to be understood to help understand the random forest mechanism. Bootstrap refers to a re-sampling procedure that randomly selects ***n*** times from ***n*** points with replacement to compute an individual tree estimate. On

the other hand, bagging refers to an aggregation scheme where many bootstrap samples are generated from the original dataset, constructed as a predictor for every sample, and aggregated to come up with a final prediction (Biau & Scornet, 2015). This implies that each successive tree is not dependent on earlier trees; hence, unstable estimates are enhanced.
The construction of individual trees is done through Classification and Regression Trees (CART)-Split criterion (Biau & Scornet, 2015). The CART-Split criterion usually is optimized at every tree node based on the impurity function to select the best cut. The Gini measure is the most commonly employed impurity function.

$$Gini = 1 - \sum_{j} P_j^2$$

*Whereby **Pj** represents the probability of class **j.***

Therefore, the algorithm for random forest techniques operates as follows. First, ***n*** bootstrap samples are drawn from the original data before constructing the trees. The second step entails building a classification tree for every bootstrap sample. A different random set of predictor variables is selected at every tree node constructed. This node determines the best-split point based on maximizing the CART-Split criterion. The last step entails predicting new data based on the aggregated forecasts from ***n*** trees (Liaw & Wiener, 2012; Biau & Scornet, 2015).

The main benefit of the random forest include the capability to estimate the importance of the predictor variables; it is a parametric classifier; hence, it does not rely on prior distribution assumptions for the sample; and it is a robust technique against overfitting, which makes it capable of handling datasets with a highly imbalanced distribution of class (Liaw & Wiener, 2012; Biau & Scornet, 2015).

2.4.2.2 Logistic Regression

Logistic regression is an algorithm where the independent and dependent variable can assume binary values, "*0*" or "*1*", or "successful" or "unsuccessful." It is the fastest and simplest algorithm in machine learning; hence, many classification problems use it as a starting point (da Silva & Bento, 2018). The essential characteristic of logistic regression is its capability to map all real numbers to the range (0, 1) (Kleinbaum & Klein, 2012). This is why it is used to transform the predictions into a number between zero and one, interpreted as a probability.

Consequently, logistic regression helps approximate the membership probability in one of the two dataset classes, as shown below.

$$P(Y = 1|\boldsymbol{X}) = \frac{1}{1+e^{-(\beta_0+\sum \beta_i X_i)}},$$
$$P(Y = 0|\boldsymbol{X}) = 1 - P(Y = 1|\boldsymbol{X}).$$

The unknown parameters, $\boldsymbol{\beta}$, are estimated based on the training dataset with data on **X**, representing a vector of input variables, and Y, representing the output or target variable. The parameter estimation method is maximum likelihood estimation (Kleinbaum & Klein, 2012).

Therefore, when values of the parameter estimate and variables of a specific early-stage start-up are inputted in the formulae, a probability of the start-up's success is obtained as the output. Such a probability represents the targeted prediction of a start-up's success. However, to assign the prediction to either of the classes, a probability threshold, ordinarily equal to 0.5, is set, as shown below.

$$1 \; if \; P(Y|X) \geq 0.5 \, ,$$
$$0 \; if \; P(Y|X) < 0.5 \, .$$

Being a linear algorithm, logistic regression assumes a linear relationship between the output and input variables (Wooldridge, 2015).

2.4.2.3 Naïve Bayes

Naïve Bayes is a conditional probability model. When the model is presented with a problem situation that needs to be classified that is denoted by a vector X =(x_1,..,x_2), which represents some ***n*** characteristics in the form of independent variables, it assigns this situation a probabilities represented by $\boldsymbol{p(C_k / (x_1,\ldots, x_2)}$, for very ***K*** possible classes or outcomes. The only challenge with this formulation is experienced when the number of characteristics ***n*** is more significant, which makes basing such a model on probability tables not feasible. Consequently, the model needs to be reformulated to ensure it is more tractable (Narasimha Murty & Susheela Devi, 2011). As a result, the conditional probability can be decomposed using Bayes' theorem, as shown below.

$$P(C_k|x) = \frac{P(C_k)*P(x|C_k)}{P(x)}$$

Naive Bayes classifier is a simplified version of a Bayes theorem-based probabilistic classifier. It's one of the rudimentary Bayesian network models, and it gets more precise results when used with kernel density estimation (Piryonesi, Madeh El-Diraby, & Tamer, 2020). Naive Bayes classifiers work under the assumption that there is no correlation between the presence or absence of two features in a class. Naive Bayes classifiers, thanks to the precision of the probability model they employ, can be trained with reasonable efficiency in a supervised learning scenario. Parameter estimation with Naive Bayes typically employs maximum likelihood.

2.4.2.4 Alternating Decision Tree (AD Tree)

An AD Tree is a classification method based on machine learning. It is associated with extending and generalizing decision trees. ADTree consists of prediction nodes containing a single number and decision nodes containing a predicate condition. ADTree categorizes an instance by traversing all paths containing true decision nodes and adding any prediction nodes traversed (Freund & Mason, 1999). ADTree contains nodes for decision and prediction. Decision nodes specify the predicate condition. Only one number is present in prediction nodes. In AD Trees, every prediction node is a root and a leaf. Numerous paths must be traversed to generate predictions, thereby dispersing the tree's knowledge (Krishna, Agrawal & Choudhary, 2016).

The sum of the prediction node values of instances that satisfy multiple splitter nodes is the aggregate prediction value. In a two-class system, a positive sum represents one class and a negative sum represents the other. Consequently, a unique interpretable tree with predictive capabilities is generated. (Krishna, Agrawal & Choudhary, 2016).

2.4.2.5 Artificial Neural Network (ANN)

From the concept of a biological neural network, an ANN is derived. It constitutes an efficient computing system composed of a large number of interconnected components arranged in some pattern. These units are known as neurons and function in tandem. Neurons are organized into interconnected layers. Multilayer perceptron (MLP) is a model of an Artificial Neural Network that is trained with a back-propagation algorithm (Depren, Aşkn, & Oz, 2017).

As shown in Figure 2.2, the multilayer perceptron consists of three components: an input layer, concealed layers, and an output layer. The weight value transports information from one neuron to another. The independent variables, which are the inputs, propagate forward using logistic activation or sigmoid function, thus, producing the dependent variables or the output values for every hidden layer. This is followed by propagating errors backward by updating the biases and weights. After that, the errors are computed, and the prejudices and weights are adjusted. These steps are repeated until such a time that the overall error is reduced to a minimum (Depren, Aşkın, & Öz, 2017).

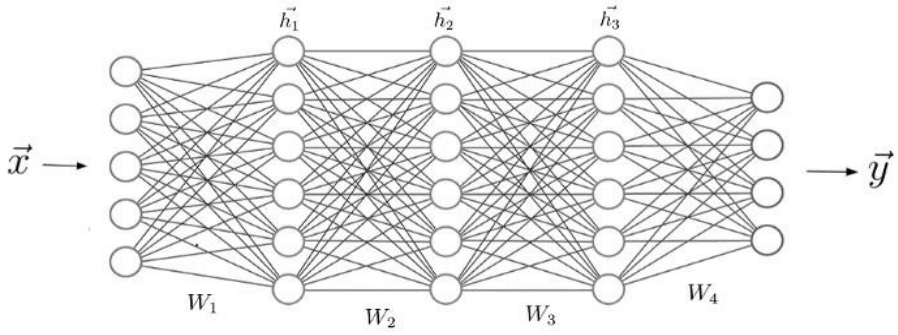


Figure 2.2: A Neural Network (Oppermann, 2019)

The initial layer of an Artificial Neural Network receives input data in the form of various texts, audio files, integers, image pixels, etc. The middle layers of an Artificial Neural Network are those that are concealed. They may consist of a single concealed layer, as in the case of perception, or multiple hidden layers. The role of the hidden layers is to perform different types of mathematical computation on the input data and recognize their patterns. The output layer, on the other hand, presents the results of the rigorous calculations performed by the hidden /middle layer (IIango & Kumar, 2017).

There are numerous parameters and hyperparameters in a neural network that affect the performance of the model. These parameters have a significant impact on the Artificial Neural Network's output. Weights, biases, learning rate, batch size, and so on are examples of these parameters. Each node in the Artificial Neural Network has a specific weight (IIango & Kumar, 2017).

2.4.2.6 Support Vector Machines

As one of the most robust prediction methods designed for binary classification, the support vector machine (SVM) Algorithm is an excellent complement to tree-based algorithms and basic logistic regression. The lack of interpretability of predictor variables is a drawback of

the SVM algorithm. The SVM algorithm "corresponds to a linear technique in a high-dimensional feature space that is nonlinearly connected to input space," even though it deals with highly complex models (Hearst et al., 1998). As a result, SVM can handle high-dimensional datasets, but the analysis is equivalent to a simple linear algorithm.

The idea behind SVM is to use a nonlinear feature function to translate the training data into a high-dimensional feature space and then build a class separating hyperplane with the most significant margin from the supporting planes (Hearst *et al*., 1998). The support planes are moved apart until they reach the initial set of observations, known as support vectors (Maroco et al., 2011). Figure 2.3 depicts the optimal hyperplane's schema. When the distance between the supporting planes is maximized, the classification objective is met, equivalent to minimizing the loss function (Maroco *et al.*, 2011).

Special kernel functions are utilized to compute separating hyperplanes, so mapping into the feature space is not explicitly carried out. All calculations are conducted directly in the input space (Hearst et al., 1998). The weighted sum of kernels evaluated at the support vector solves the classification issue (Maroco et al., 2011).

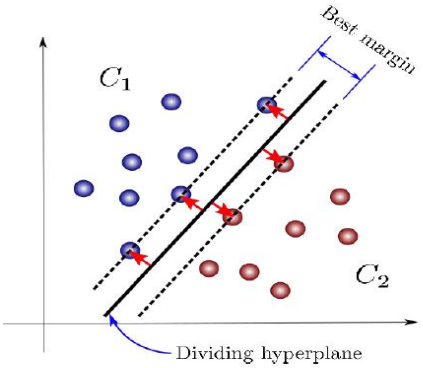


Figure 2.3: Support Vector Machines Schema (Carrasco, 2019)

2.5 Models and Frameworks

A machine learning framework is a tool that allows data scientists and software developers to create machine learning models without having to work out the underlying mathematical and statistical principles of the machine learning algorithms. It simplifies the development process by eliminating the need for programmers to reinvent the wheel when creating a

specific application. As discussed below, machine learning frameworks include several similar working libraries that make developing machine learning models easier.

2.5.1 Frameworks

2.5.1.1 TensorFlow

TensorFlow is a flexible and scalable deep learning framework used in various applications (Abadi et al., 2017). Dataflow graphs are used in TensorFlow to describe computation, shared state, and mutating operations. It partitions a dataflow graph's nodes among several machines in a cluster and numerous computational devices within a single machine, such as multiple cores on a single CPU, multiple graphics processing units, and specialized ASICs called Tensor Processing Units (TPUs). In contrast to earlier "parameter server" systems, which incorporated shared state management into the system, TensorFlow gives programmers the freedom to try out new approaches to optimization and training. While TensorFlow may be used for many other things, deep neural network training and inference are where it shines.

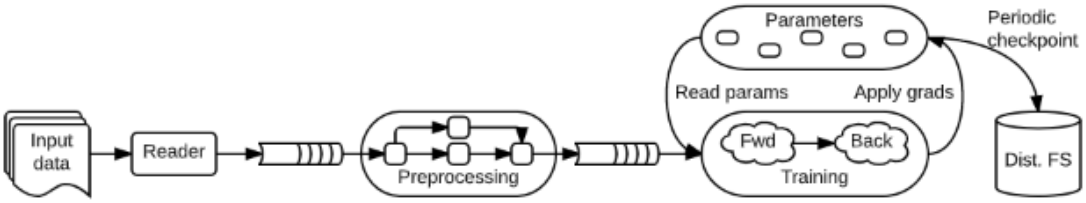


Figure 2.4: A schematic TensorFlow dataflow graph for a training pipeline (Abadi et al., 2017)

2.5.1.2 PyTorch

Deep learning systems have historically had to trade off efficiency for usability. PyTorch is an example of a machine learning framework that successfully demonstrates that these two aims can coexist by providing support for an imperative and Pythonic programming style that supports code as a model, simplifies debugging, and is consistent with other popular scientific computing tools. It's effective, and it works with GPUs and other hardware accelerators. (Paszke et al., 2019).

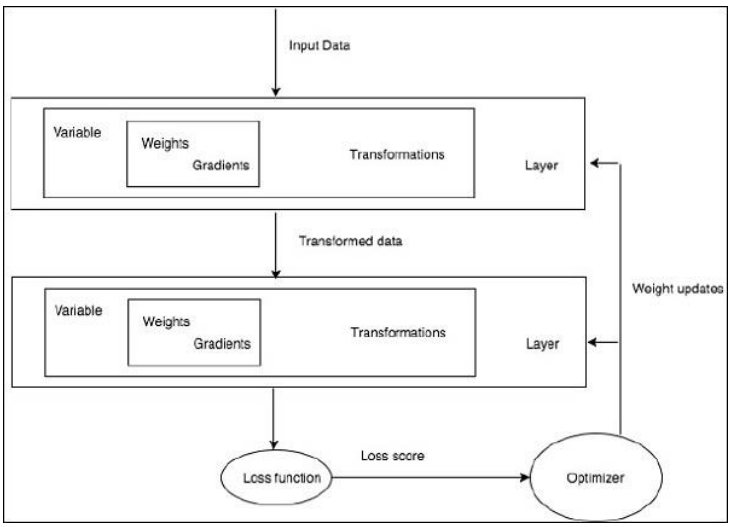


Figure 2.5: Schematic overview of PyTorch Framework (Paszke et al., 2019)

2.5.1.3 Keras

Keras is a Python wrapper framework that provides bindings to deep learning libraries including TensorFlow, CNTK, Theano, and the recently announced Deeplearning4j (Keras 2018). It was designed for fast testing and is accessible for no cost under the MIT license. Keras's underlying frameworks make it possible for it to run equally well on both graphics processing units (GPUs) and central processing units (CPUs) (Nguyen et al., 2019).

2.5.1.4 Scikit Learn

Scikit-Learn is a well-known Python open-source tool renowned for its extensive library of DM/ML algorithms. (Scikit 2018). David Cournapeau founded Scikit-Learn while participating in the Google Summer of Code program. It extends the capabilities of NumPy and SciPy with data mining methods for classification, clustering, regression, dimensionality reduction, preprocessing, and model selection. Graph-based software Matplotlib is also utilized. (Nguyen et al., 2019).

Data for use with any objects or algorithms from scikit-learn can be provided as two-dimensional arrays of size-sampled features. This convention gives it a broad scope and makes it applicable to any discipline. Estimators can apply models to data, Predictors can make predictions based on new data, and Transformers can transform data from one

representation to another; all Scikit-learn objects share this consistent set of operations (Abraham et al., 2014). Below is a rundown of the three Scikit Learn objects:

i) Estimator

The estimator interface provides access to a suitable method for automatically determining appropriate values for model parameters based on available training data. Classification, regression, and clustering are supervised learning algorithms that can be found as objects implementing this interface. Feature selection and dimensionality reduction are two examples of machine learning tasks that can be used as estimators (Abraham et al., 2014).

ii) Predictor

A predictor is an estimator based on the predict technique, which makes predictions for each sample in an input array X. (Abraham et al., 2014). The fact that it can be applied to unanticipated input led to its rebranding as the "X test." In supervised learning, the estimated model is used to make predictions, and this method often outputs those estimates as labels or values.

iii) Transformer

Certain estimators (called transformers) employ a transformed approach to preparing data for a learning algorithm better. The preprocessing, feature selection and dimensionality reduction method transformers are all included in the library's collection of tools. The term "inverse transform" refers to a technique that can be used if the transformation can be reversed.

2.5.2 Models

2.5.2.1 Hybrid Intelligence

Dellermann et al (2017) proposed a hybrid intelligence model to predict early-stage business start-up success. They argued that machines could not correctly interpret "soft" or unqualifiable information. Therefore, it would be essential to have a hybrid system that leverages both human intuition and machine capabilities. The model accepts raw data and processes it following the stages detailed in Figure 2.6. The first process is the conversion of the signals into readable format. Automation is then performed using Artificial Neural Networks, Naïve Bayes, Logistic Regression, Random Forest and Support Vector Machine. The last process involves a judgment task, namely aggregation.

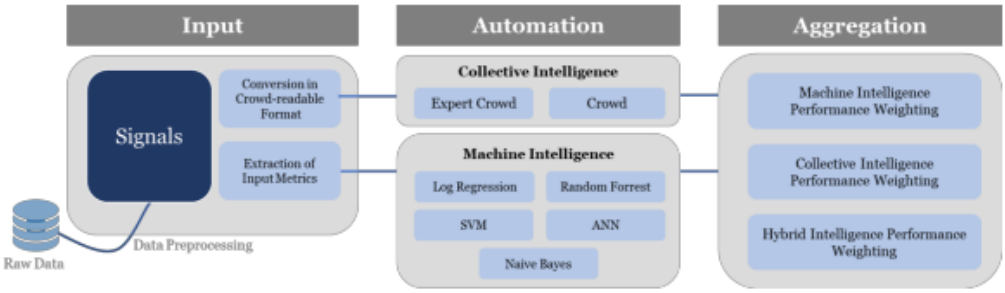


Figure 2.6: Hybrid Intelligence Model (Dellermann et al., 2017)

2.5.2.2 Random Forest and Support Vector Machine (SVM)

A random forest model and SVM were proposed by Li (2021). This framework prioritizes the most critical aspects of a startup's environment, such as the company's willingness to accept venture financing, the sort of market in which it operates, and the city in which it is based. Rounds of funding are also an essential metric, with rounds A and B money being particularly significant.

Figure 2.7: Random Forest (Li, 2020)

2.5.2.3 Predictive Models

Krishna et al (2016) proposed a predictive model that combined more than 30 classification schemes and narrowed it down to six. The six chosen were Naïve Bayes, AD Trees, BayesNet, Lazylbl, Random Forest, and Simple Logistics. The WEKA toolkit was used for classification analysis and modelling. The flow of the predictive models is shown in Figure 2.8.

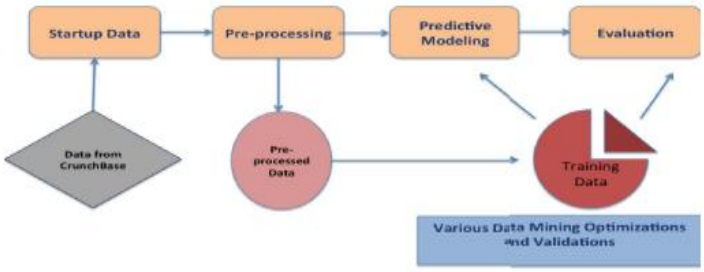


Figure 2.8: Overview of the predictive models (Krishna, Agrawal & Choudhary, 2016)

2.6 Gaps in the Existing Systems

Despite recent advancements in startup success prediction, a review of existing systems has highlighted several deficiencies in the methods used. These shortcomings have limited the effectiveness and applicability of the current machine learning (ML) approaches in dealing with the problem at hand. Firstly, the existing literature primarily focuses on predicting the success rates of established firms rather than start-up firms. This narrow focus renders the current models ill-suited for accurately predicting the success of start-ups due to the significant disparities between corporate and start-up success prediction. Start-ups operate under unique conditions and face distinct challenges compared to established companies, such as limited resources, high uncertainty, and rapid growth potential. Consequently, the existing ML methods fail to capture the specific factors that influence the success or failure of start-up ventures.

Moreover, most of the studies in the literature rely on data collected at a single time point, disregarding the temporal aspect of failure. Start-ups' trajectories and outcomes evolve over time, making it crucial to consider longitudinal data to gain a deeper understanding of failure triggers. By incorporating panel data that capture percentage or growth metrics, such as changes in employee count, funding growth rates, and other relevant factors tracked over time, the accuracy of prediction models can be significantly improved. These dynamic metrics provide valuable insights into the developmental patterns and trends that are essential for more accurate predictions.

Additionally, the existing literature often lacks customization to specific industries and their subcategories. Different industries, particularly disruptive ones like cryptocurrency-focused digital and tech organizations, have unique success criteria and distinct factors driving their success compared to traditional sectors like utilities or heavy machinery. Current ML methods do not adequately account for industry-specific variations, resulting in limited accuracy when applied to diverse sectors. Customizing the quantitative models to suit the specific requirements of various industries would enable the identification of industry-specific success drivers, ultimately leading to more precise predictions of start-up success within each sector.

Lastly, there is a significant gap in the availability of user-friendly tools for entrepreneurs and investors without technical expertise to predict the success of start-ups. Many existing ML methods require a deep understanding of technical concepts and expertise in data analysis, making them inaccessible to individuals without specialized skills. As a result, entrepreneurs and investors may struggle to make informed decisions regarding start-up investments. Developing user-friendly prediction tools that simplify complex ML algorithms and provide intuitive interfaces would empower non-technical users to assess the potential success of start-ups more effectively.

2.7 Conceptual Model

The conceptual framework is the rationalization of the problem at hand and a proposed way the researcher intends to solve the problem. A user will input the start-up features through the user interface. The system queries the start-up success from the compiled Scikit Learn model and then returns a decision to the user. The model was trained and compiled using Artificial Neural Networks (ANNs). The data used to train the model was obtained from Crunchbase and merged to join across the columns with different surface forms. The consumer-facing interface allows entrepreneurs and investors to input the various features. Based on the independent variables, an inference is made, and a prediction is given on whether the start-up will succeed. Figure 2.9 depicts the conceptual model of the tool.

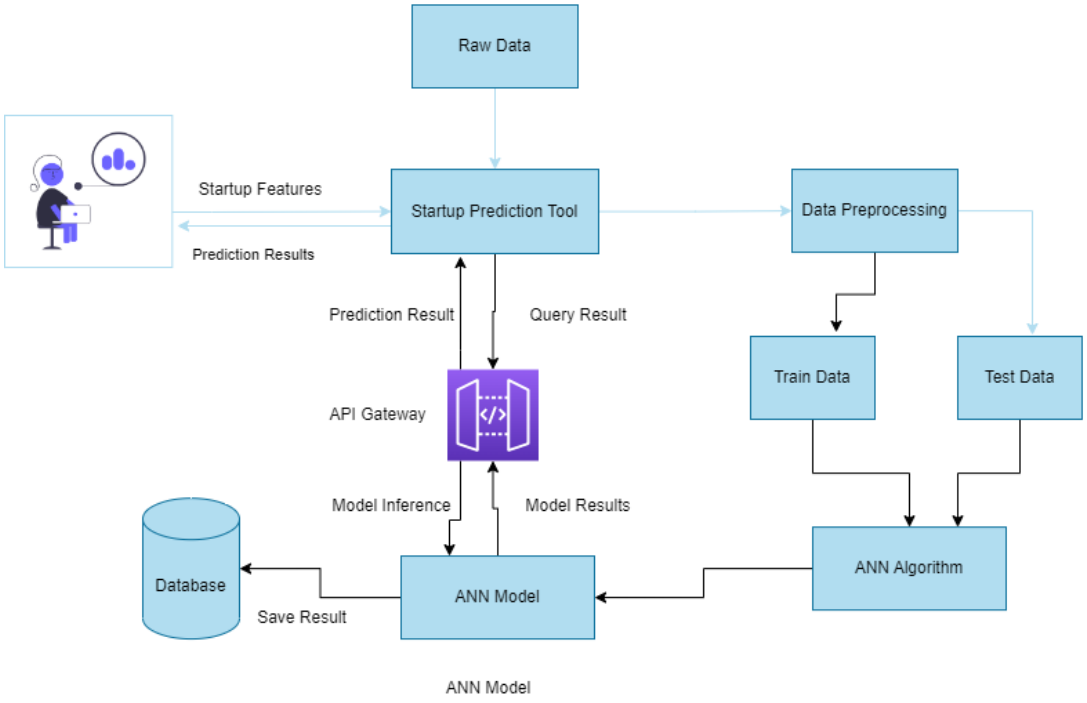


Figure 2.9: Conceptual Model of the solution

Chapter 3: Research Methodology

3.1 Introduction

This chapter details the research strategy employed to determine the data features that influence start-up success prediction, the method used to select the most appropriate machine learning model, and the process followed to implement a prediction model for start-ups in their formative stages. Tandon (2010) defines research methodology as "a systematic approach to solving research problems that specify the numerous processes followed by the researcher in examining a research subject and the logic behind them."

The choices adopted regarding research design, data collection, and the target population provide the researcher with legitimacy and perspective on the study's boundaries. The methodology used supports the research's validity (Somekh & Lewin, 2005).

3.2 Research Design

A research design is a comprehensive plan for conducting a study so that all of its parts work together logically and consistently to answer the research question. It can be used as a template for research and analysis (Mishra & Alok, 2011). Both descriptive and practical research methods were used in this investigation. Descriptive research designs collect data to describe phenomena, situations, or populations comprehensively. It specifically aids in addressing the what, when, where, and how questions concerning the research challenges rather than the why (Mishra & Alok, 2011).

Therefore, the descriptive research design helped gather information about early-stage start-ups' success factors. On the other hand, applied research is a scientific study that aims to answer practical problems (Hedrick, Bickman, & Rog, 2013). This was appropriate for the study since it seeks to solve the challenge of predicting start-up success or failure.

3.3 Target Population

The targeted population was start-ups funded in Africa for the past six years, from 2015 to 2021. Thousands of start-ups have emerged from 2015 to 2021; some are thriving, while others are shutting down offices. The profiles obtained from CrunchBase informed the population of the study. Figure 3.1 shows the distribution of the start-ups in selected African countries in 2020 alone.

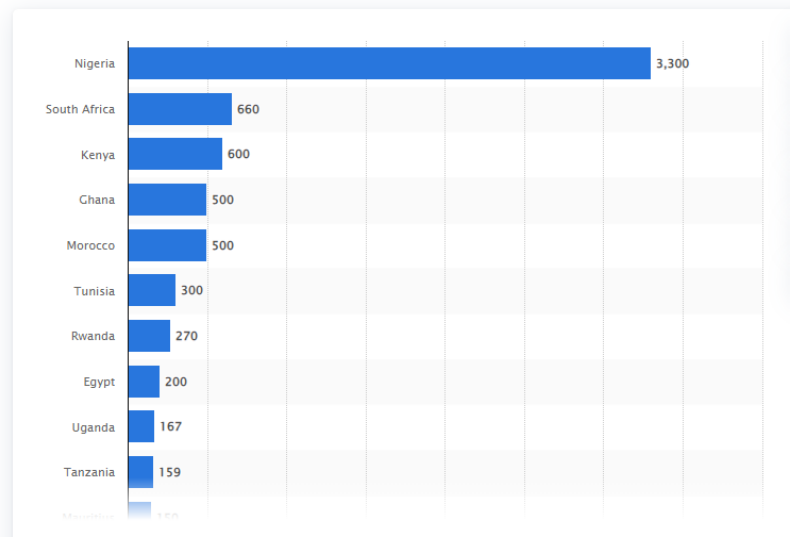


Figure 3.1: Number of tech start-ups that raised funds in Africa in 2020 alone (Statista, 2020)

3.4 Sample Size

A sample refers to a part of a population from which inferences can be drawn about the entire population. A descriptive survey requires a sample size of 10% to 30% (Mugenda & Mugenda, 2003). However, since the study sought to establish a model that can predict the success of early start-ups, where more data was required, 100% of the entire dataset was used.

To prevent the model from becoming biased, which would give a misleading appearance of greater model accuracy, 80% of the dataset was used for training, and 20% was used for testing. The 80/20 rule is widely employed in machine learning because it provides a reasonable balance between training the model and assessing its effectiveness using the available data. Using 80% of the data as the training set and 20% as the testing set, allows the model to be trained on a significant enough sample to capture the patterns in the data while also providing an estimate of its generalization performance on new data (Dobbin & Simon, 2011).

3.5 Data Collection

The study used data obtained via an Application Program Interface (API) from the CrunchBase website. CrunchBase is an online database of information about businesses big and small, public and private, worldwide. Details regarding funding and investments,

founding members and key personnel, acquisitions, mergers, market trends, and recent news are all examples. The collection includes details about a wide range of international startups. Date of establishment, place of establishment, the sum of all funding rounds, valuation, market worth, investments, and acquisitions. Despite its humble beginnings, CrunchBase has become an indispensable tool for numerous SMB-focused venture capital, consulting, marketing, sales firms, and academic research projects.

CrunchBase offers the data needed for this study to accurately forecast the chances of a start-up's success. The target variables are the number of financings rounds a start-up can attract, the attributes of an entrepreneur, and mergers and acquisitions. The features (target variables) were grouped into general, target, and investor features. CrunchBase serves as a repository for all start-up and investor data, thus suitable for the chosen features for study. The second part focused on factors determining a start-up's success. The third section focused on start-up prediction attributes. Entrepreneurs/Investors (Users) are required to input the information through a start-up's evaluation form, which returns a result based on the features provided.

3.6 Research Quality and Reliability

Criterion validity was employed in this research. The degree to which a criterion or concrete validity measure relates to an outcome (Taherdoost, 2016). It determines how accurately one metric can predict the performance of another. The validity will be chosen because the research is geared towards predicting success, which will be done using several measures. Cronbach's Alpha is the dependability test that will be performed. The reliability of a survey's battery of questions can be evaluated using Cronbach's alpha (Frost, 2022).

3.7 System Development Methodology

The Agile Development Systems Methodology, along with the OOAD (Object-Oriented Analysis and Design) software life cycle, was employed in the development of the start-up success prediction tool. Agile development encompasses various iterative and incremental software development methodologies such as Scrum, Crystal, and Lean development. This methodology was specifically chosen for this research due to its iterative nature, allowing for continuous feedback, refinement, and delivery of the software system.

The OOAD software life cycle was integrated into the agile development process to provide a structured approach to system analysis, design, and implementation. It involves several phases, including requirements gathering, analysis, system design, implementation, and

testing. These phases help ensure a systematic and comprehensive approach to software development, aligning with the agile methodology's iterative and incremental nature.

The development process consists of continuous planning, design, building, testing, review, and launch iterations, known as sprints. This iterative approach allows for regular feedback and collaboration between the development team and stakeholders, facilitating continuous improvement and adaptation throughout the software development life cycle.

The Manifesto for Agile Software Development articulates the following four tenets of agile development:

i). Individuals and interactions are prioritized over procedures and instruments.
ii). Functional software over exhaustive documentation.
iii). Collaboration with customers over contract negotiation; and
iv). adapting to change over sticking to a plan.

Figure 3.2 provides an overview of the agile methodology.

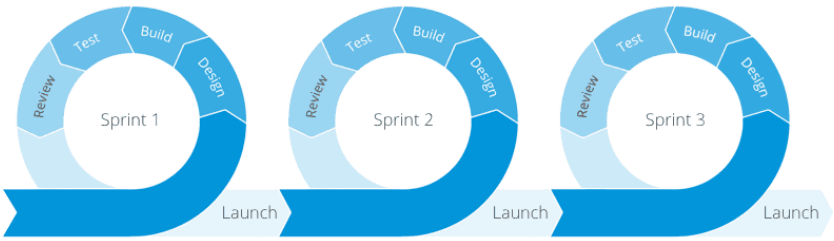


Figure 3.2: Overview of Agile methodology (Moniruzzaman & Hossain, 2013)

3.8 Start-up Success System Evaluation

The system's evaluation used both the prediction accuracy results and the usability-based approach to assessing expert systems. The testing set, precision, and recall are all available in the Keras, and Scikit Learn python libraries. Also, F1-score and accuracy are utilized to evaluate classification methods. The rate at which classification is correct is known as its accuracy. The F1-score is the harmonic mean of the recall and precision scores. Precision estimates the proportion of correctly identified samples among positive samples, while recall

estimates the proportion of correctly identified samples across all positive samples. Mathematically, they are presented as shown below.

$$\textbf{Accuracy} = (TP + TN)/(TP + TN + FP + FN)$$

$$\textbf{Precision} = TP / (TP + FP)$$

$$\textbf{Recall} = TP / (TP + FN)$$

$$\boldsymbol{F1} = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

Whereby,

True positive (TP) is a result for which the prediction model correctly predicted the positive class.

True Negative (TN) denotes a result for which the prediction model accurately predicted the negative class.

False Positive (FP) = an outcome where the prediction model mispredicts the positive class.

False Negative (FN) = a result in which the prediction model mispredicts the negative class.

A confusion matrix helps describe a classification model's performance, as shown in Table 3.1.

Table 3.1: Confusion Matrix

	0 (Predicted Negative)	1 (Predicted Positive)
0 (Actual Negative)	TN	FP
1 (Actual Positive)	FN	TP

3.9 Utilization and Dissemination of Research Results

The success of business start-ups in their early stage is essential to any economy. The research results are expected to aid investors and other stakeholders in the start-up scene in measuring the likelihood of their business success at its early stage. This will help reduce the failure rate witnessed in the current environment as businesses will pivot early, and investors will not lose millions of money investing in a business destined to fail. The

research will primarily be disseminated in the existing open-access public repository of the university.

3.10 Ethical Considerations / Issues

The researcher will ensure the confidentiality of the data obtained from reliable databases and datasets. The core dataset source for this is Crunchbase, and permission to use the dataset has been obtained from the source. Since the research touches on an essential part of the business ecosystem, it will ensure the reliability of the outcomes.

Chapter 4: System Analysis and Design

4.1 Introduction

Systems analysis involves gathering requirements using both quantitative and qualitative measures. Techniques for understanding the current system and identifying new requirements (traits or capabilities that the system must have in order to satisfy the users' needs) include: talking to the system users, observing the systems in action, and collecting and analyzing data (Conger & Mason, 2013). This chapter delves into the system architecture and the general design concerning the requirements raised by the stakeholders and the potential users of the early-stage start-up prediction tool. The analysis informs the requirements of the system (both functional and non-functional requirements). Unified Modelling Language(UML) diagrams have been used to explain the system's architecture and demonstrate the interaction between the target users and the system's main components.

4.2 Requirement Specifications

During the requirement gathering process, it was of utmost importance to ensure that the process was unbiased and comprehensive. In order to achieve this, multiple approaches were taken to gather the necessary information for the development of the startup prediction tool. One approach involved observing various start-up prediction tools in action. This firsthand experience provided valuable insights into the challenges and limitations associated with existing tools. By directly observing the tools in use, it was possible to identify areas that needed improvement and determine the specific functional requirements necessary for an effective startup prediction tool.

Additionally, the requirement gathering process included a thorough review of existing literature. Several research papers and studies were examined, such as Makarenko et al. (2019), Okrah, Nepp, and Agbozo (2018), Murimi (2014), and Aktan (2011). These sources provided valuable insights into the functional requirements that were essential for accurately predicting the success of startups. By utilizing the findings from these sources, the researcher ensured that the prediction tool would encompass all the necessary features and capabilities to meet the identified requirements.

By combining the firsthand observation of existing tools with the insights gathered from the existing literature, the requirement gathering process was designed to be unbiased and comprehensive. This approach aimed to ensure that the developed startup prediction tool

would effectively address the needs and challenges faced by startups, providing accurate predictions of success based on relevant and reliable functional requirements.

The requirement gathering process was driven by a commitment to impartiality and a thorough understanding of the functional requirements essential for a successful startup prediction tool. The combination of firsthand observations and a comprehensive review of existing literature ensured that the development process was grounded in a broad range of perspectives and knowledge, resulting in a robust and unbiased set of functional requirements for the tool.

4.2.1 Functional Requirements

The tool should allow the user to:

i). Create an account.
ii). Login.
iii). View their profile.
iv). Enter the features of a start-up for evaluation.
v). View the results of the prediction.
vi). Predict success or failure with high accuracy.

4.2.2 Non-Functional Requirements

Non-Functional requirements define the tool's quality attributes. The Non-Functional requirements are as follows:

i). The tool should be easy to use.
ii). The tool should be secure against attacks.
iii). The tool should be available whenever the user needs to use it.
iv). The tool should be easy to maintain.
v). The tool should be fault tolerant.
vi). The tool should be compatible with various operating systems and browsers.

4.3 System Architecture

The system's architecture consists of three main components: the web interface, the machine learning model, and the database. The web interface provides a point of interaction between the user and the model. The database stores the users' prediction results for future reference whenever they need to. The system users must input the start-up features to the interface as

prompted. The features are then sent to the machine learning model for prediction purposes. The predictions' results are stored in the database and displayed to the user, as illustrated in Figure 4.1.

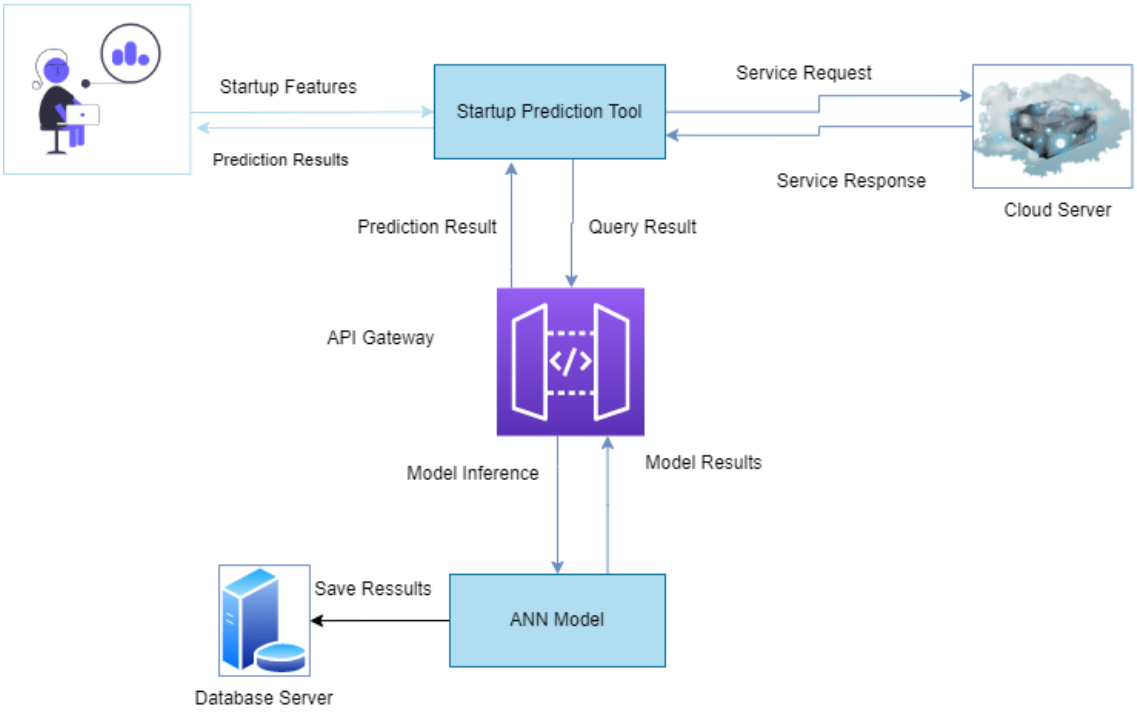


Figure 4.1: System Architecture

4.4 System Design

Systems design, a fundamental part of the development of any information system, is the procedure of outlining the system's structure, components, modules, interfaces, and data to fulfill its intended purposes (Rouhani & Lecic, 2018). Object-oriented analysis and design is the approach of choice for this study's design (OOAD). The Object-Oriented Analysis and Design (OOAD) technique is a software engineering approach to designing software systems by creating object-oriented models that abstract critical elements of the intended system and the subsequent use of the models to direct development. The model's principles and notation depict choices made during system design that will significantly affect the finished product (Carstoiu & Grigorescu, 1995).

4.4.1 Use Case Diagram

Use Case Diagrams to summarize how a user interacts with the system. Users can evaluate the system behaviour using Use Case Diagrams before writing code. Similarly, they can be used as a blueprint throughout software development (Mule & Waykar, 2015). The Use Case diagram for the start-up success prediction tool is shown in Figure 4.2.

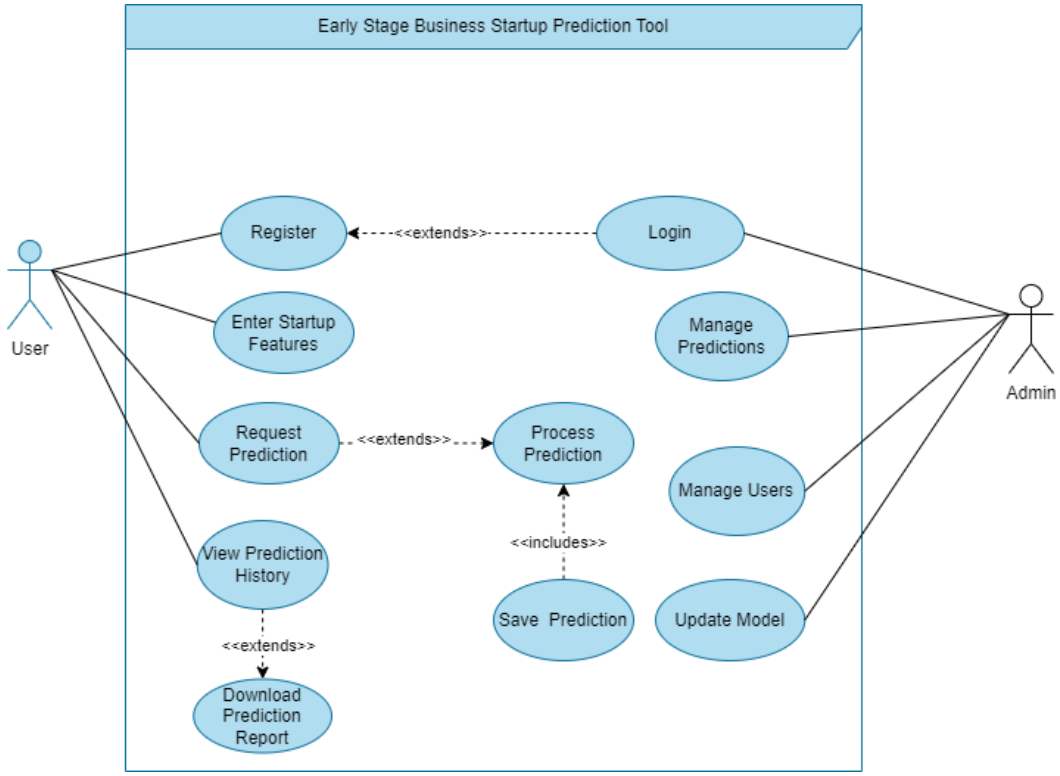


Figure 4.2: Use Case Diagram

4.4.1.1 Detailed Use Case Descriptions

Table 4.1 shows the detailed description of use cases earlier presented in Figure 4.2

Table 4.1: Description of use cases

Use Case	Pre Conditions	Main Success Scenario	Post Conditions
Login	The user is registered on the system	-User enters login credentials -User logs in to the system and is directed to the home page.	None
Register	The user is connected to the internet.	-User enters their details on the registration form -User registration details are saved in the database.	None
Enter Start-up Features	The user is logged in	-User completes the start-up evaluation form. -Start-up Evaluation is initiated after the user presses evaluate button	Start-up evaluation result
View Profile	The user is logged in	-User clicks on the view profile button. -User profile is displayed	None
View Prediction Results	Start-up Features are keyed into the system	-User Clicks on View Prediction Results -Prediction results displayed to the user.	None
View Prediction History	User is authenticated	-Prediction history displayed	None
View Prediction Report	User is authenticated	-User clicks on the View Report button. -Prediction Report is presented to the user.	None
Admin Login	None	-Admin enters login credentials into the login form. -Admin logins successfully.	None
Manage Users	Admin is logged in	-Admin selects a user. -Admin edits, updates, or removes users.	
Manage Predictions	Admin is logged in	-Admin selects a prediction. -Admin views, edits, updates, and deletes prediction successfully.	

4.4.2 Class Diagram

One of the most helpful types of UML diagrams is the class diagram, which accurately depicts a system's structure by modelling its classes, characteristics, processes, and object relationships, as shown in Figure 4.3 (Bergstrom et al., 2022).

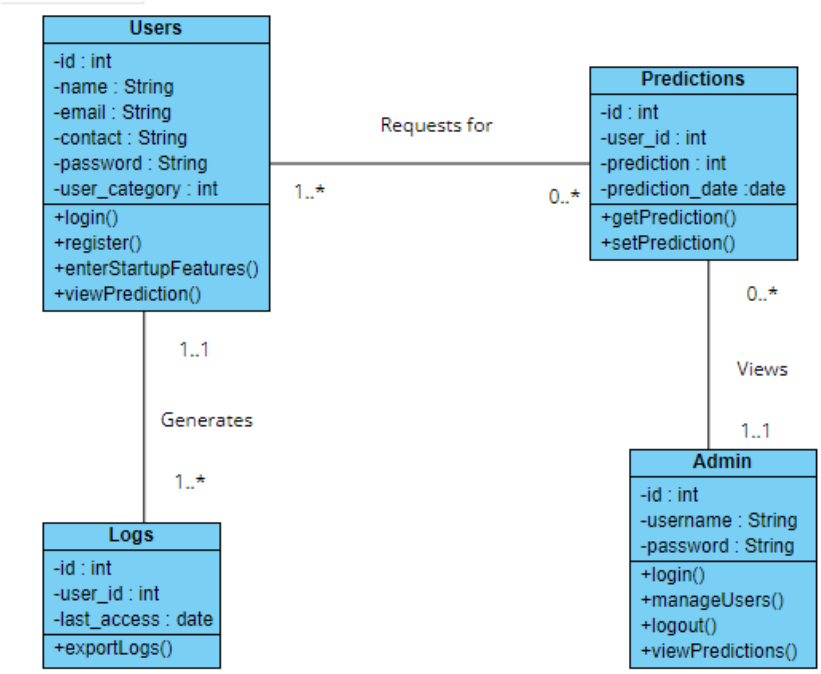


Figure 4.3 Class Diagram

4.4.3 Sequence Diagram

A sequence diagram is an interaction diagram showing the causes and effects of a chain of events (Alvin et al., 2019). The tool has five objects which interact with each other starting from the user, web interface, and ANN model (comprising of the setup, training, and model). Figure 4.4 illustrates how the objects interact and the feedback mechanisms.

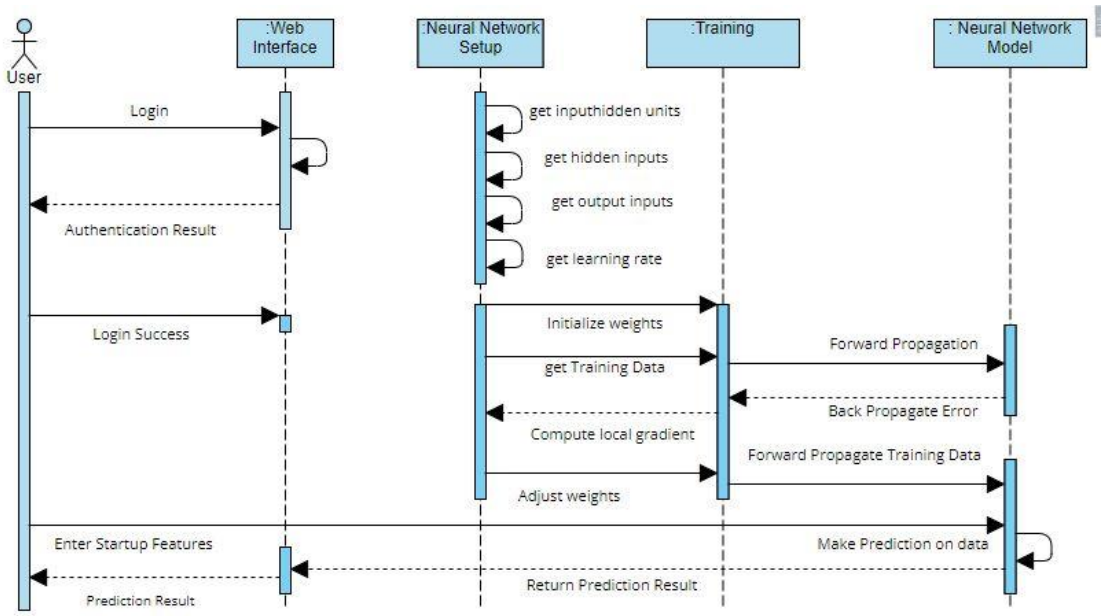


Figure 4.4: Sequence Diagram

4.4.4 Database Schema

A database schema is a blueprint for a database that outlines how different tables and models interact. The tool's database structure is depicted in Figure 4.5.

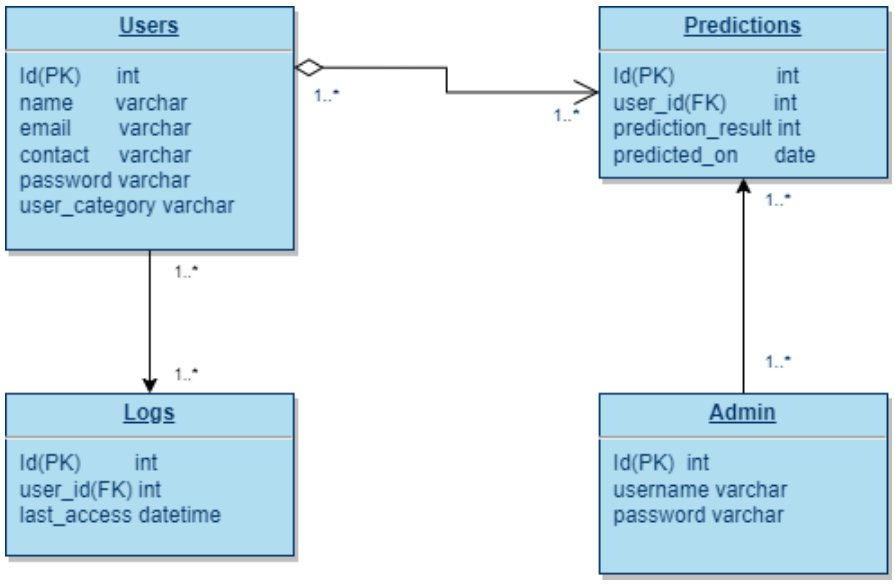


Figure 4.5: Database Schema

4.5 Wireframes

A system wireframe is a design used to inform the implementation of a system by rendering key intended features of the system based on user requirements or the developer's perspective. It is a representation of how the system functions. However, the system may provide more advanced features than the highlighted ones or features that fall short of the design (de Lange et al., 2020).

4.5.1 Home Page Wireframe

Figure 4.6 shows the landing page wireframe. This will be the first page the user will see once they visit the web application on their browser.

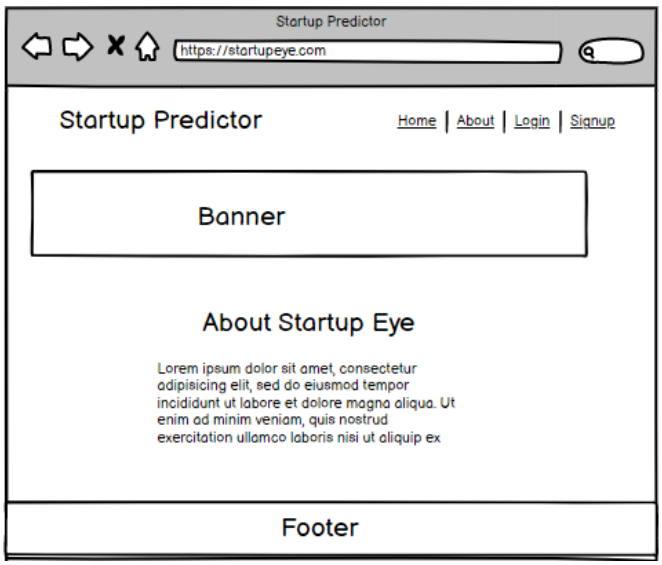


Figure 4.6: Home Page Wireframe

4.5.2 Login Wireframe

The wireframe, as shown in Figure 4.7, contains a form that requires one to login with their correct credentials, namely Email, and Password. After entering the credentials, there is an option to click the Login button.

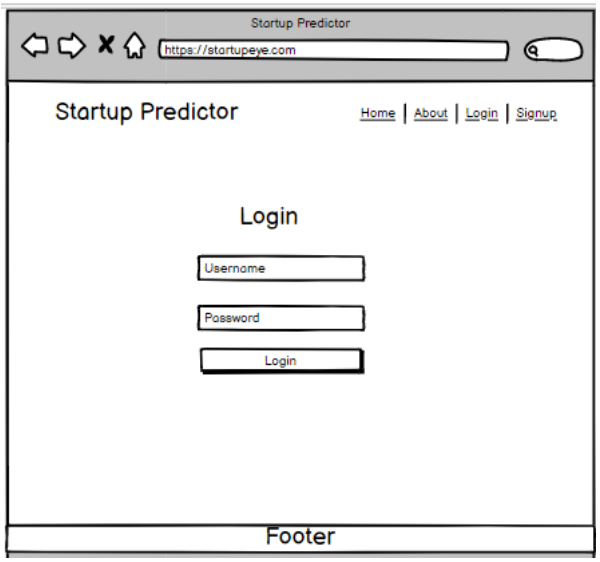


Figure 4.7: Login Wireframe

4.5.3 Register Wireframe

The wireframe, as shown in Figure 4.8, contains a form that requires one to create an account by providing the requested information.

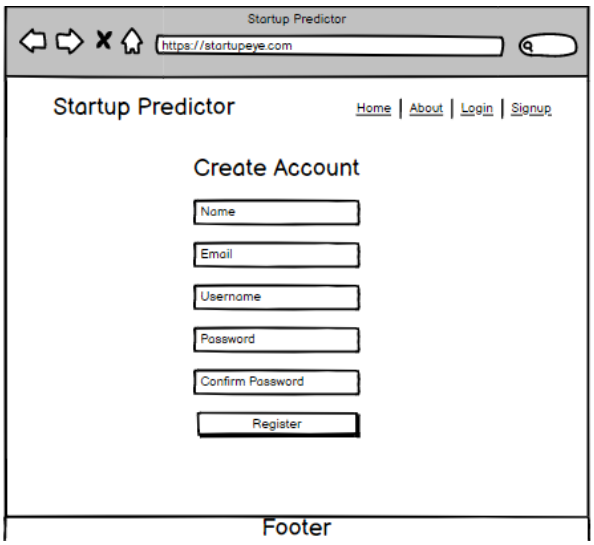


Figure 4.8: Register Wireframe

4.5.4 Start-up Evaluation Page Wireframe

The start-up evaluation form captures details of a start-up and allows the prediction to be performed after. Figure 4.9 shows the start-up evaluation wireframe.

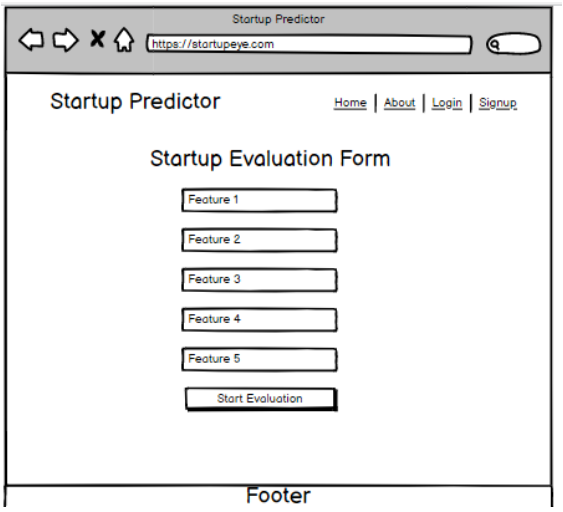


Figure 4.9: Start-up Evaluation Form Wireframe

4.5.5 Results Page Wireframe

This wireframe illustrated in Figure 4.10, shows how the user feedback will be provided after requesting predictions from the system. Users will view the detailed description of the results and status of whether a start-up will succeed or fail.

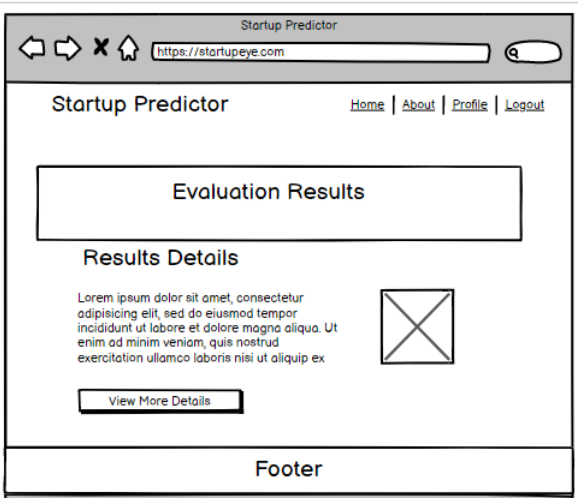


Figure 4.10: Prediction Results Wireframe

4.5.6 History Wireframe

This wireframe shows the history of user predictions on the system. A user can see the date and time of the prediction and the results. Figure 4.11 depicts the history wireframe.

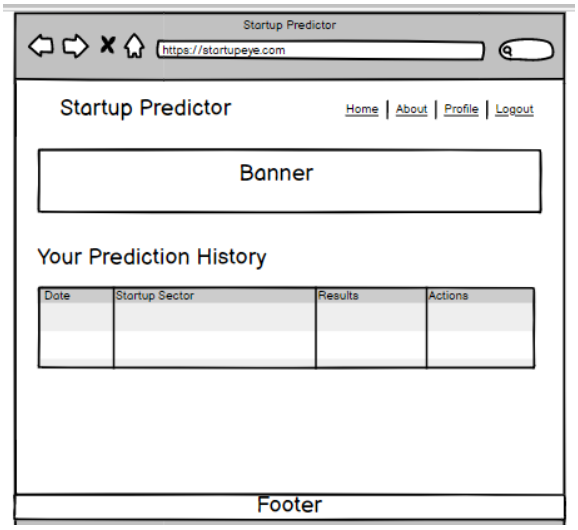


Figure 4.11: Prediction History Wireframe

Chapter 5: System Implementation and Testing

5.1 Introduction

This section focused on the system's development, testing, and certification. The implementation phase involved analyzing the system's modules, development process, and operation. The testing and validation process consisted of functional and usability testing to determine whether the system met its objectives.

5.2 Model Components

5.2.1 Artificial Neural Network (ANN) Layers

The neural network model was created using Keras. The model is sequential, which indicates that layers are added consecutively. The first layer is fully connected (also known as a dense layer) with 64 nodes, and it accepts shape as input (*X train.shape[1]*). This indicates that *X train.shape[1]* characteristics must be included in the model's input. Rectified linear unit (*ReLU*), a typical activation function in neural networks, is utilized at this layer. Only for positive values does the *ReLU* activation function apply a linear function to the input. For negative values, it yields zero.

The second layer is also a fully connected 32-node layer with *ReLU* activation mechanism. This layer receives its input from the first layer's output. Also, the third layer is an entirely interconnected layer with 16 nodes and a *ReLU* activation mechanism. This layer receives its input from the second layer's output. The fourth and final layer is interconnected with eight nodes and a *ReLU* activation mechanism. This layer receives its input from the third layer's output. Figure 5.1 and 5.2 shows the code snippet and model architecture consisting of neural network layers.

```
# define the model architecture
model = Sequential()
model.add(Dense(64, input_shape=(X_train.shape[1],), activation='relu'))
model.add(Dense(32, activation='relu'))
model.add(Dense(16, activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(y_train.shape[1], activation='softmax'))
```

Figure 5.1: Model Architecture Code

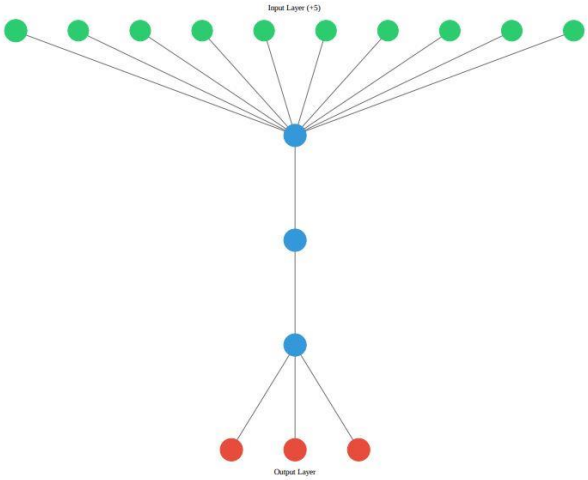


Figure 5.2: Model Architecture

5.3 Web Application Components

The application has three primary interfaces: a homepage, a start-up evaluation interface, and a forecast result interface. Using Tailwind CSS, the front-end application's user experience was designed to be current and aesthetically pleasing. The model's application program interface was interacted with using a PHP backend. Detailed descriptions of the components are as follows:

5.3.1 Home Page

The homepage provides a summary of the application and its function. It is intended to entice and urge users to utilize the system. The homepage contains a description and a call-to-action button that goes to the interface for evaluating the start-up.

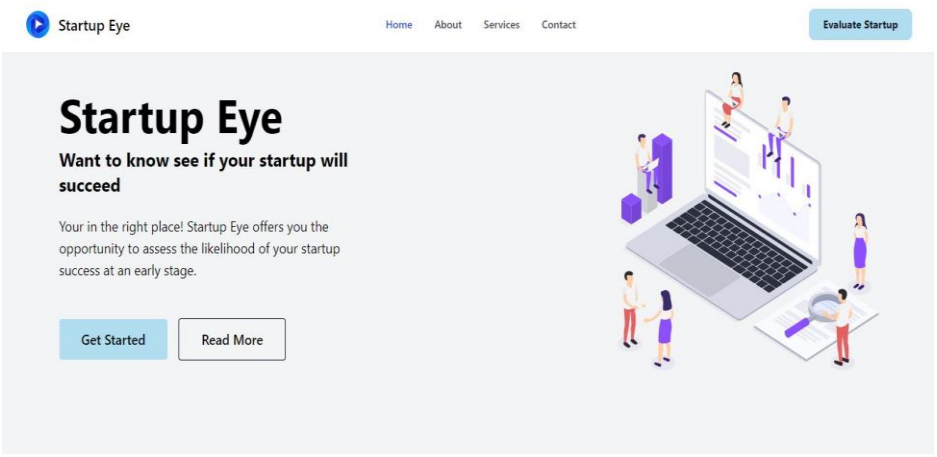


Figure 5.3: Home Page

5.3.2 Prediction Interface

This interface is the main feature of the application. It allows users to evaluate their start-up's potential for success by completing the form. The form has fields that require users to provide features such as start-up industry, funding rounds, entrepreneurs' characteristics, etc. Users can enter their answers into a form, which is then used to predict their start-up's success. The design of the interface is user-friendly, with instructions and feedback mechanisms that are both straightforward. Figures 5.4 and 5.5 show the start-up evaluation form.

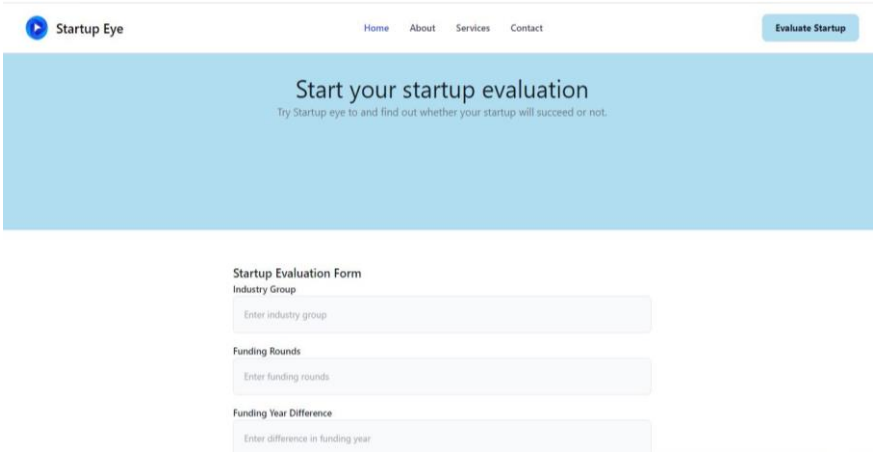


Figure 5.4: Evaluation Form

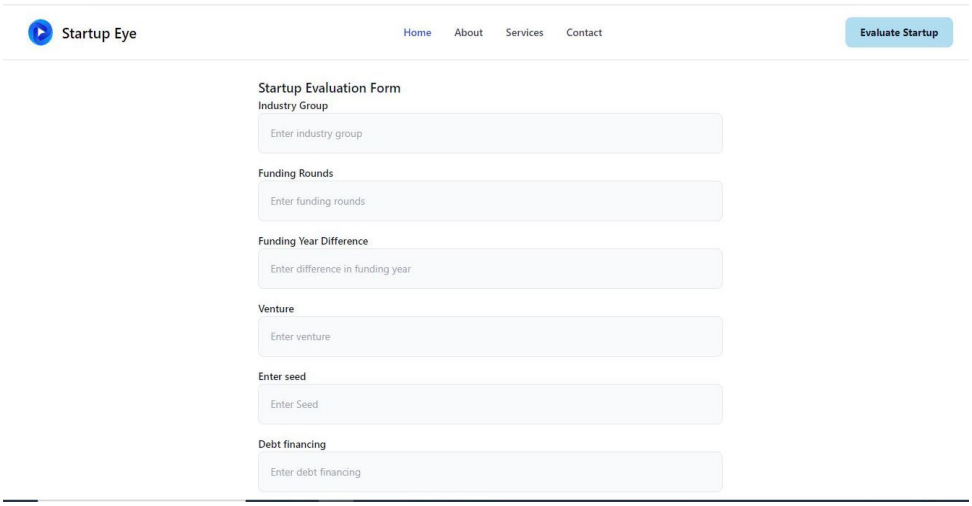


Figure 5.5: Evaluation Form

5.3.3 Prediction Results

Once the user has completed the start-up evaluation form, the prediction result interface displays the predicted success or failure. Figure 5.6 illustrates a screenshot of the prediction results.

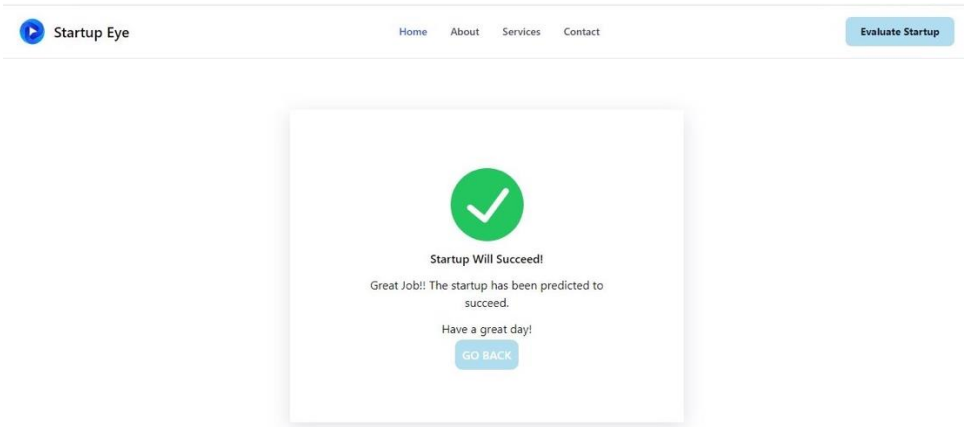


Figure 5.6: Predicted Result

5.4 System Implementation

During this study and the tool development phase, Agile software development methodology was used. Its capacity for continuous iteration facilitated the modification of various system versions to satisfy the study's objectives. The start-up prediction prototype is a powerful solution that employs cutting-edge machine learning and deep learning technologies to analyze and predict the success of start-ups.

The tool is designed to help investors and business owners assess a start-up's viability before devoting major resources to it. The tool accurately predicts a start-up's likelihood of success by examining historical trends, industry standards, and other pertinent data. The Flask framework integrates the machine learning model into the user interface to create a smooth user experience. Users can submit pertinent information about their firm, such as its industry, funding, funding rounds, amount of funds raised, etc., and receive a comprehensive analysis of the start-up's potential. In addition, the tool presents recommendations for areas where enhancements should be made to boost the probability of success. The start-up prediction tool is an innovative solution that utilizes cutting-edge technologies to deliver essential insights into the start-up industry. Its user-friendly design and sophisticated algorithms make

it an indispensable tool for investors and business owners who wish to make data-driven business decisions.

5.4.1 Development Environment

The following software and hardware environment was used to create the tool:

i. Python
ii. Mac Operating System
iii. VS Code
iv. Google Colab
v. Scikit-Learn
vi. Keras
vii. PHP (Frontend Development)
viii. Flask (API Development)
ix. Tailwind CSS

5.4.2 Start-up Dataset Collection

The CrunchBase website was utilized to obtain the startup dataset, which consisted of over a million entries. CrunchBase serves as a comprehensive database containing valuable information about public and private businesses worldwide. It offers detailed insights into various aspects of startups, including funding and investments, key personnel, acquisitions, mergers, current trends, and news.

The dataset of African startup data consisted of 11,215 entries, representing a significant number of startups from various countries across the African continent. This dataset provides a valuable resource for studying and understanding the dynamics of the African startup ecosystem. The size of the dataset ensures a robust and comprehensive analysis, allowing for a more accurate understanding of the startup ecosystem in Africa. This extensive collection of data serves as a valuable foundation for conducting further research, developing strategies, and making informed decisions to support the growth and success of African startups.

The dataset obtained from CrunchBase encompassed a wide range of startup businesses, providing essential details such as start date, location, initial funding, total funding rounds, valuation, market value, investment, and acquisition. With its extensive coverage and relevance to venture capital, consulting, marketing, sales firms, and academic research on

small and medium-sized businesses, CrunchBase proved to be an indispensable resource for this study.

5.4.3 Start-up Data Pre-processing

In the start-up data pre-processing phase, the collected data from CrunchBase underwent a series of cleaning, filtering, and processing steps to ensure its suitability for analysis. Firstly, the start-up data obtained from CrunchBase was filtered to include only African start-ups, narrowing down the dataset to focus on African countries. The subsequent step involved data cleaning, which aimed to address missing or erroneous data to ensure the accuracy and completeness of the dataset. To handle missing values, the data was examined for NaN (Not a Number) entries, and rows containing NaN values were dropped using the dropna() function in the Pandas library. This process ensured that the dataset only consisted of complete and usable data.

Additionally, the relevant columns were transformed into floating-point numbers to enable mathematical operations and calculations on the data. Specifically, the funding total column, initially read as an object type, was cleaned and converted into a numerical column using the astype() function in Pandas. This conversion allowed for meaningful calculations such as computing means, medians, and standard deviations. The cleaning procedures described above, including dropping rows with NaN values, removing spaces from column fronts, and converting the funding total column to a numerical format, were carried out to prepare the data for further analysis. A visual representation of these data pre-processing processes can be observed in Figures 5.7, 5.8, 5.9, and 5.10.

```
# Read the file using the detected encoding
df = parse_file("/content/investments_VC.csv")
```

Figure 0.7: Load CSV file

```
# Dropping rows where NaN is present
df = df.dropna(subset=['permalink', 'name', 'homepage_url'])
df
```

Figure 0.8: Cleaning of data

```
cols = ['funding_rounds', 'funding_total_usd', 'seed', 'venture', 'equity_crowdfunding',
       'undisclosed', 'convertible_note', 'debt_financing', 'angel', 'grant',
       'private_equity', 'post_ipo_equity', 'post_ipo_debt',
       'secondary_market', 'product_crowdfunding', 'round_A', 'round_B',
       'round_C', 'round_D', 'round_E', 'round_F', 'round_G', 'round_H']
df.loc[:,cols] = df.loc[:,cols].astype(int)
```

Figure 0.9: Converting Columns to Floating Point

```
df3 = df2[['cat_status', 'cat_Industry_Group'
       ,'cat_funding_rounds',
       'cat_diff_funding_year', 'cat_total_investment' ,
       'cat_equity_crowdfunding', 'cat_venture', 'cat_seed', 'cat_undisclosed',
       'cat_convertible_note', 'cat_debt_financing', 'cat_angel', 'cat_grant',
       'cat_private_equity', 'cat_post_ipo_equity', 'cat_post_ipo_debt',
       'cat_secondary_market', 'cat_product_crowdfunding', 'cat_round_A',
       'cat_round_B', 'cat_round_C', 'cat_round_D', 'cat_round_E',
       'cat_round_F', 'cat_round_G', 'cat_round_H']]
```

Figure 5.10: Selecting Columns for the Model

5.4.4 Exploratory Analysis

After performing the data cleaning process, an exploratory analysis was conducted on the startup dataset to gain insights into the available information. One aspect explored was the distribution of companies based on their status types. The dataset revealed that there were 848 acquired companies, 529 closed companies, and a significant majority of 9528 operating companies. This information provided an understanding of the current landscape of the startup ecosystem. Figure 5.11 shows the distribution of startups based on their status.

status	count	mean	std	min	25%	50%	75%	max
acquired	848.0	2.000000	1.463447	1.0	1.0	1.0	3.0	15.0
closed	529.0	1.523629	1.114602	1.0	1.0	1.0	2.0	9.0
operating	9528.0	1.722817	1.342359	1.0	1.0	1.0	2.0	14.0

Figure 5.11 Distribution of start-ups

Funding rounds for the various status was also explored. According to the analysis, there were a total of 848 acquired companies with an average of two funding rounds. The standard deviation of funding rounds for acquired companies was approximately 1.46, indicating some variability in the number of rounds. The minimum and maximum number of funding

rounds for acquired companies were 1 and 3, respectively. The range of funding rounds for acquired companies suggests that they typically had between 1 and 3 rounds, with an average of 2.

For closed companies, there were 529 in total, with an average of 1.52 funding rounds. The standard deviation of funding rounds for closed companies was approximately 1.11, indicating less variability compared to acquired companies. The minimum and maximum number of funding rounds for closed companies were both 1. Thus, closed companies typically had 1 funding round, on average.

In the case of operating companies, the dataset included 9,528 entities, with an average of 1.72 funding rounds. The standard deviation of funding rounds for operating companies was approximately 1.34, suggesting a moderate degree of variability. The minimum and maximum number of funding rounds for operating companies were 1 and 14, respectively. This indicates that operating companies had a wider range of funding rounds, with an average of around 1.7.

The analysis also focused on the characteristics of the startups in terms of their country codes and markets. The dataset contained 47 unique country codes, indicating the diverse geographical representation of the startups. Moreover, there were 573 unique markets represented among the startups. Figure 5.12 and 5.13 shows the unique codes and unique markets.

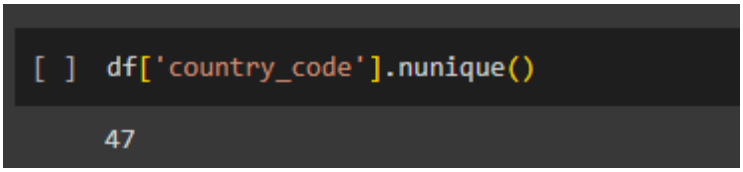


Figure 5.12 Unique Country Codes

Figure 5.13 Unique Markets

Further exploration revealed the top five markets based on the amount of funding they received. Biotechnology secured the highest funding followed by Mobile, Software, Clean Technology and Healthcare. These findings highlighted the prominent industries and markets within the startup ecosystem.

Additionally, the analysis examined the funding durations of the companies. It was observed that there were relatively few companies with a funding duration exceeding 13 years, indicating that most companies received funding within a shorter time frame. This insight provided an understanding of the typical funding dynamics within the dataset.

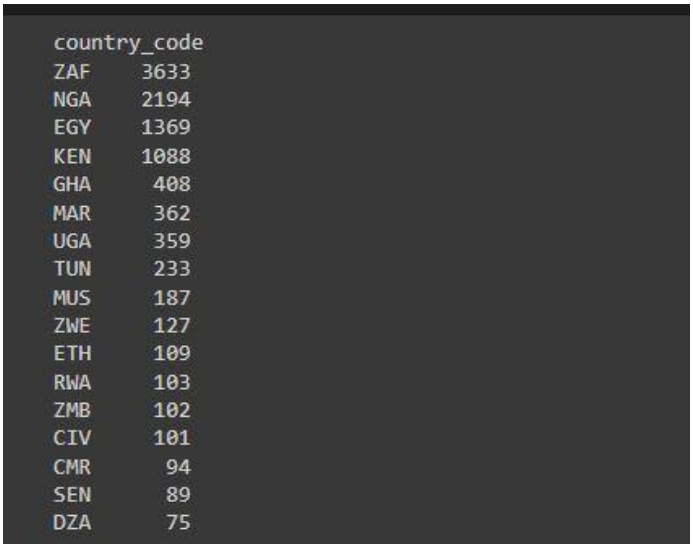
The analysis also included examining the industry groups represented in the dataset, the distribution of companies across these industry groups, and the number of companies in each country code. The dataset consisted of 42 different industry groups. Among these, Software emerged as the most prominent industry group, with 2,046 companies identified within this category. Other notable industry groups included Biotechnology with 1,087 companies, Healthcare with 742 companies, Internet Services with 615 companies, and Commerce and Shopping with 553 companies. This analysis provided a comprehensive understanding of the diverse range of industry groups present in the startup ecosystem as shown in Figure 5.14.

Industry_Group	
Software	2046
Biotechnology	1087
Health Care	742
Internet Services	615
Commerce and Shopping	553
Information Technology	454
Mobile	452
Advertising	440
Community and Lifestyle	343
Financial Services	327
Media and Entertainment	284
Gaming	239
Professional Services	230
Data and Analytics	225
Education	217
Other	188

Figure 5.14 Industry Groups

Furthermore, the analysis focused on the number of companies in each country code. The findings indicated that South Africa had the highest number of startups, as most of the

companies in the dataset were from this country. This insight highlighted the significance of South Africa as a hub for entrepreneurial activity within the African startup landscape.



country_code	
ZAF	3633
NGA	2194
EGY	1369
KEN	1088
GHA	408
MAR	362
UGA	359
TUN	233
MUS	187
ZWE	127
ETH	109
RWA	103
ZMB	102
CIV	101
CMR	94
SEN	89
DZA	75

Figure 5.15 Number of Companies in each country

To further refine the analysis, variable selection was performed using the Recursive Feature Elimination (RFE) method. The top 10 features were selected based on their importance in predicting the outcome variable. This process involved training a model and ranking the features based on their significance. The selected variables included cat_Industry_Group, cat_funding_rounds, cat_diff_funding_year, cat_total_investment, cat_venture, cat_seed, cat_round_A, cat_round_B, cat_round_C, and cat_round_D. These features were identified as the most influential in predicting the desired outcome and would serve as valuable inputs for subsequent analyses and modelling tasks.

5.4.5 Training Model

The dataset is divided using the train-test-split approach from the scikit-learn library in the first stage. X stores the input features, whereas Y stores the target variable. The test size is set to 0.2, which reserves 20% of the data for evaluating the model's performance. The parameter for the random state is set to 42 for reproducibility. The training data is then transformed into pandas dataframe with the original X column identifiers.

The next step is data normalization, performed on the training dataset's numerical columns. *MinMaxScaler* method from the scikit-learn library is used to scale the features from 0 to 1. This is necessary to ensure that the features are on the same scale and have a similar impact

on the model's predictions. The test data is also transformed using the scaler fitted to the training data.

Data encoding is then performed on the target variable, which is a categorical column. *to_categorical* method from *keras.utils* library converts the target variable to a one-hot encoded format. This ensures that a unique binary vector represents each category. The same transformation is applied to the test data.

The model architecture is defined using Keras Sequential API in the next step. The model consists of 5 layers of densely connected neurons. The first layer comprises 64 neurons that receive the training dataset's input shape. This layer's activation function is rectified linear unit (ReLU), a popular choice for deep learning models. The subsequent layers have 32, 16, and 8 neurons using the *ReLU* activation function, as observed in Figure 5.19. The final output layer contains the same number of neurons as categories in the target variable and employs the softmax activation function to generate the probability distribution over the categories.

The model is then compiled using the compile method of Keras. The optimizer used is Adam, with a learning rate of 1e-4, a famous optimizer for deep learning models. The loss function used is *categorical_crossentropy*, which is appropriate for multi-class classification problems. The model's performance is evaluated using an accuracy metric.

The model is then trained using the Keras fit technique. Input parameters include the training dataset, the target variable, the number of epochs, and the group size. One hundred epochs are used to train the model, containing 32 data points for each block. The training procedure involves adjusting the model's weights to minimize the loss function. The trained model is then capable of predicting the categories of new startups.

```
# split the dataset into training and testing sets
X = df4.drop("cat_status", axis=1)
y = df4["cat_status"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
X_train = pd.DataFrame(X_train, columns = X.columns)

"""## 2. Data Normalization"""

# normalize the numerical columns
scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

"""## 3. Data Encoding"""

# one hot encode the categorical columns
y_train = to_categorical(y_train)
y_test = to_categorical(y_test)

"""## 4. Model Definition"""

# define the model architecture
model = Sequential()
model.add(Dense(64, input_shape=(X_train.shape[1],), activation='relu'))
model.add(Dense(32, activation='relu'))
model.add(Dense(16, activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(y_train.shape[1], activation='softmax'))
```

Figure 5.16: Model Training

5.4.6 Creating Model API

The API defines a function *predict_startup_success* which takes in input features related to a startup, such as industry group, funding rounds, total investment, etc., and returns a prediction of either "Success" or "Failure" based on the input features. The input features are pre-processed using the *MinMaxScaler* and *StandardScale*r methods from the scikit-learn library to ensure consistency in the range and distribution of the input data. The prediction is made using a pre-trained machine learning model loaded using the *CustomUnpickler* module.

The API endpoint /predict accepts POST requests with input data as a JSON payload. The payload contains input features such as industry groups, funding rounds, etc. The endpoint uses the *predict_startup_success* function to forecast the start-up's likelihood of success or failure. and returns the result as a JSON response, as shown in Figure 5.20.

```
@app.route('/predict', methods=['GET', 'POST'])
def startup_success():

    predicted_success = "Success"
    if request.method == 'POST':

        industry_group = request.form['industry_group']

        funding_rounds = request.form['funding_rounds']
        diff_funding_year = request.form['diff_funding_year']
        total_investment = request.form['total_investment']
        venture = request.form['venture']
        seed = request.form['seed']
        debt_financing = request.form['debt_financing']
        angel = request.form['angel']
        private_equity = request.form['private_equity']
        round_A = request.form['round_A']
        round_B = request.form['round_B']
        round_C = request.form['round_C']
        round_D = request.form['round_D']
        round_E = request.form['round_E']
        round_F = request.form['round_F']

        predicted_success = predict_startup_success(industry_group, funding_rounds

    return jsonify({'Prediction ' : predicted_success})
```

Figure 5.17: API

5.4.7 Start-up Prediction Tool

The software can only be used with an internet connection and a web browser. Data from Crunchbase was used to train the model to predict start-up success using an artificial neural network.

5.5 System Testing

The web-based tool was created utilizing the Agile development methodology, which permits continuous iterations throughout various phases. Performance and problem issues were examined with continuous testing. The developed prototype was subjected to functional testing to determine if it met the specified functional requirements.

5.5.1 Test on Model Accuracy

Several metrics, including precision, recall, and F1 score, were used to evaluate the model's precision and performance. These measures would help to determine the model's capacity to correctly distinguish between start-ups with a high likelihood of success and those that

are unlikely to succeed. The model was tested using 20% of the dataset, achieving a maximum accuracy of 0.8681. This is represented in Figure 5.21.

```
# evaluate the model on the testing data
test_loss, test_acc = model.evaluate(X_test, y_test)
print("Test Loss: ", test_loss)
print("Test Accuracy: ", test_acc)
```

```
64/64 [==============================] - 0s 4ms/step - loss: 0.4625 - accuracy: 0.8681
Test Loss:  0.46245482563972473
Test Accuracy:  0.8681209683418274
```

Figure 0.18: Model Evaluation

5.5.2 System Validation

The system validation involved 120 participants who tested the start-up prediction system and provided feedback. Most participants found the system easy to use and user-friendly and recommended it to others. This is a positive result, indicating that the tool has met the needs of its users and is likely to be successful.

5.5.2.1 Usability Testing Results

A usability validation questionnaire was made using Google Forms and given to 140 respondents via email and social media. Everyone was able to respond to the question. As shown in Figure 5.22, all participants had access to the web application.

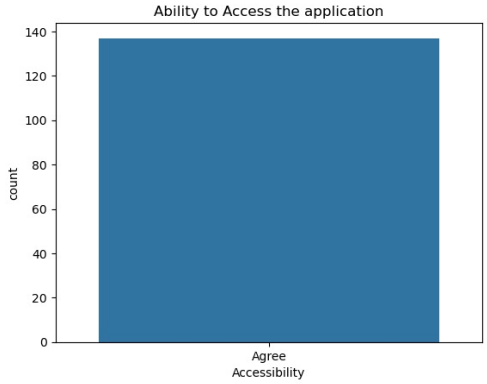


Figure 0.19: Usability results

None of the participants encountered problems when requesting startup prediction, as shown in Figure 5.23.

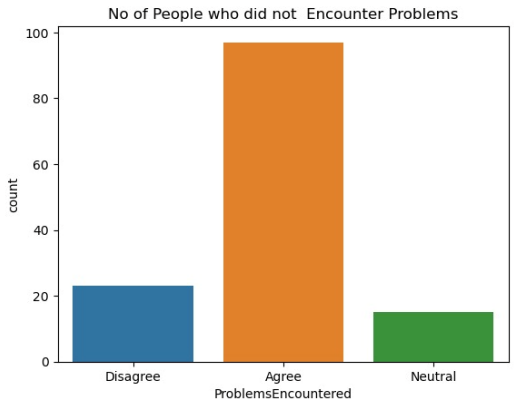


Figure 0.20: Problems encountered

All the participants confirmed they would recommend the tool to other users to help predict start-up success.

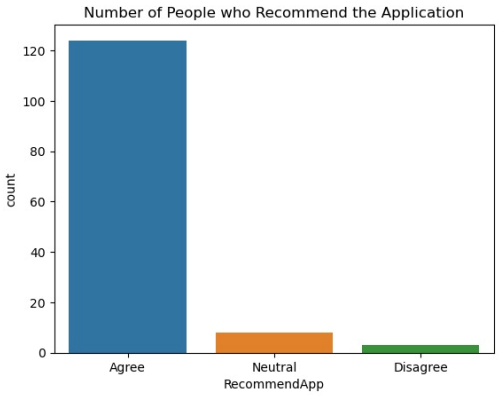


Figure 0.21: Tool acceptability

5.6 Conclusions

During the implementation phase, the needs that emerged from the analysis acted as a guide. The system design detailed how system development was conducted. The research objectives served as the foundation for the system's development.

Chapter 6: Discussions

6.1 Review of Research Objectives

According to the findings, limited access to capital, limited market opportunities, and regulatory and legal barriers are some of the factors impeding the success of start-ups. These results are consistent with the literature review described in section 2.2.

The second objective was to examine the methodologies used to forecast the success of new businesses. The most relevant technology to create a viable solution was chosen using the research findings. The literature review examined the various technologies that support the prediction of startup success. Żbikowski and Antosiuk (2021) employed three methods to predict start-up success: Support Vector Machines (SVM), Logistic Regression, and XGBoost. The researchers obtained data on 213,171 companies from Crunchbase and based their success indicators on the organizations' acquisition, IPO, or continued operation while receiving Series B capital. The researchers achieved 86% accuracy with Logistic Regression, 84% with SVM, and 86% with XGBoost. The accuracy levels of their study are comparable to the findings of our study, despite the different amounts of data used for training and testing.

Aktan (2011) sought to forecast start-up success using a sample of 180 enterprises in the production industry; the financial ratios were derived from annual reports. Five machine learning methods were examined by the author: K-Nearest Neighbor, Support Vector Machines, Decision Trees, Bayesian models, and Artificial Neural Networks. This prediction was made for well-established companies with sufficient data for success prediction.

Krishna et al (2016) created a prediction model for start-ups based on the critical events in the various lifecycles of a new firm. The researchers proposed a method for predicting the outcome of a start-up based on a number of critical factors, such as the amount and duration of seed funding, the timing of the Series A investment, and factors influencing the company's success or failure at each milestone. Their research heavily influenced the selection of variables included in this study. All of these factors were crucial in assessing whether a start-up will succeed.

Using the pre-processed data, the researchers employed a variety of data mining classification approaches, as well as data mining optimizations and validations. Techniques

such as Random Forest, ADTrees, Bayesian Networks, etc., were utilized to offer them analysis. The established system lacks specific capabilities. Initially, the accuracy levels were poor and varied between the various development methods, averaging 87.6 per cent. In addition, the model was not usable by non-technical individuals, who made up the most significant proportion of the intended audience. However, the accuracy obtained in their study is closer to 86.87% achieved in this study.

Li's (2010) research produced a methodology for categorizing start-ups and analyzing the critical success factors for new businesses. The researcher utilized Random Forest and Support Vector Machine to study and explain some of the critical factors that impact the success of start-up businesses. Kaggle provided the researcher with data. Using CrunchBase, the data had information regarding 22,000 start-ups launched between 1997 and 2014. Random Forest and SVM algorithms averaged 88.5% accuracy and 98.5% precision. The algorithms used in his study achieved better accuracy and precision scores than those obtained in this study. This can be attributed to the amount of data used for training since his research was double the amount compared to this study.

The third objective was to examine the existing prediction algorithms and models to predict business startup activity. The research results were used to identify the most suitable algorithms and select the best models. The literature review examined the algorithms utilized by various models to predict startup success. The study found that machine learning techniques utilized in startup prediction play a crucial role in classification, analysis, association, and prediction duties. The utilized technology, the Artificial Neural Network algorithm, is consistent with the tasks performed by the technologies discussed in section 2.4.

The fourth objective was to develop a machine learning tool to forecast the success of African startups in their early stages. Due to the inherent uncertainty of launching a business, investors desired a tool to assist them in predicting the success or failure of start-ups, according to the research findings. By predicting whether a startup will be successful, the devised system can assist investors in determining which startups to fund.

The fifth objective was to validate and evaluate the tool. Usability testing was conducted to evaluate the application's performance, and none of the participants encountered any issues while interacting with the system. Core functionality accessibility, responsiveness, and utility were rated as outstanding.

6.2 Advantages of the Tool

The developed tool helps entrepreneurs and investors predict the success of their start-ups at an early stage. This guides them to make informed decisions when deciding on whether to invest in a start-up or pivot respectively.

6.3 Limitations of the Tool

i. Training a machine learning model to predict the success of a business requires an extensive and diverse dataset encompassing the features of successful and failed start-ups. In the case of Africa, however, the dataset is limited in size and information publicly available.
ii. The application is only suitable for start-ups and not any business in the latter stage of development.

Chapter 7: Conclusion and Recommendation

7.1 Conclusions

In conclusion, the development of a prediction tool for early-stage companies utilizing artificial neural networks (ANN) has demonstrated significant promise for evaluating the feasibility of African firms. By restricting Crunchbase data to African businesses only, the model attained an 86% accuracy rate in forecasting the success of early-stage start-ups. The tool can deliver insightful information to investors and entrepreneurs in Africa's rapidly expanding start-up ecosystem.

In recent years, considerable interest has been in using artificial intelligence (AI) to forecast start-up success. The high failure rate of new initiatives has been a significant issue for investors and business owners, making it imperative to establish methods for evaluating the feasibility of new ventures. The ANN model for African start-ups assessed various characteristics, including the business strategy, management team, industry, and funding. The incredible accuracy of the model indicates the feasibility of applying machine learning algorithms to predict start-up success.

The capacity of the developed ANN model to evaluate big datasets and uncover patterns that may be difficult for humans to perceive is one of its main advantages. Discovering features and trends unique to the African start-up ecosystem was possible by training the model on data from African start-ups. This method is beneficial when dealing with complicated datasets, making the model an indispensable instrument for investors and business owners who wish to examine many African start-ups.

The tool may also assist investors and entrepreneurs in identifying potential hazards linked with African start-up companies. By analyzing the dataset, the model can uncover potential flaws in a firm's business model or management team, thereby aiding investors and entrepreneurs in making informed judgments. Also, the tool can assist entrepreneurs in identifying areas of their start-up that need work, improving their likelihood of success.

In addition, the technology can facilitate access to finance for African businesses by giving investors more precise assessments of their feasibility. Due to a lack of investment opportunities and the idea that they are high-risk endeavours, it is sometimes tricky for African start-ups to obtain financing. This perception can be altered by the established

methodology, which provides investors with more precise assessments of the potential success of African entrepreneurs.

Traditional techniques of predicting company success, which mainly rely on subjective assessments by investors and entrepreneurs, are significantly less accurate than the model's 86% accuracy rate. The proposed approach can make more objective and data-driven forecasts using AI and machine learning algorithms. This strategy can assist in lessening the risk associated with investing in early-stage entrepreneurs, which can substantially influence the African economy.

However, it is essential to emphasize that the created model is not an ideal answer for predicting start-up success. Like any machine learning method, the model is only as accurate as the training data. Consequently, verifying that the training data represents the African start-up environment is essential. In addition, the model's accuracy may decline if new start-up ecosystem patterns arise in Africa.

Forecasting the success of early-stage start-ups is a complex and challenging undertaking that involves a mix of numerous elements and strategies. In this dissertation, we have investigated some of the most significant indicators of start-up success, such as founder traits, market size, and funding patterns, and examined the strengths and limits of various prediction models. The findings indicate that there is no one-size-fits-all strategy for predicting the success of early-stage start-ups and that a combination of qualitative and quantitative methodologies is required for effective forecasting.

The development of an ANN-based prediction tool for early-stage businesses in Africa is a crucial step toward evaluating the viability of new ventures in Africa. The technology has the potential to give investors and entrepreneurs valuable insights, enabling them to make more educated decisions regarding which firms to invest in. With the African start-up ecosystem expanding fast, the tool is a timely and much-required solution to the continent's high start-up failure rate. The insights gathered from this study can give investors, entrepreneurs, and policymakers valuable counsel for fostering the growth of early-stage firms. By harnessing the most recent research and innovations in start-up prediction, we can ensure that potential enterprises receive the resources and support they require to flourish and contribute significantly to the economy and society.

7.2 Recommendations

This study showed that Artificial Neural Networks could be used in a start-up prediction system. The researcher recommends addressing the influence of social and cultural elements on start-up success. This may involve improving diversity and representation in the start-up ecosystem, helping underrepresented groups, and investigating the impact of networks and social capital on the success of start-ups.

7.3 Future Work

The researcher saw that the tool could be expanded in the future. To further enhance user experience, a mobile application could be developed for Android and IOS users. It would also be interesting if the future application could provide recommendations on improvement areas once the start-up has been evaluated.

Future research might examine the possible impact of government policies and initiatives on the forecast of start-ups in their infancy stages. By studying the efficacy of policies such as tax incentives, grants, and regulatory frameworks, we may be able to develop better strategies to assist the growth and success of start-ups, especially in new industries and locations.

References

Abadi, M., Isard, M., & Murray, D. G. (2017). A computational model for TensorFlow: an introduction. Proceedings of the 1st ACM SIGPLAN International Workshop on Machine Learning and Programming Languages. https://doi.org/10.1145/3088525.3088527

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*. https://doi.org/10.3389/fninf.2014.00014

Achleitner, A. K. (2010, May 20). Start-up-Unternehmen. Retrieved from Gabler business dictionary: https://wirtschaftslexikon.gabler.de/definition/start-unternehmen-42136/version-145394

Achtenhagen, L., Brundin, E., & Springerlink (Online Service. (2016). *Entrepreneurship and SME Management Across Africa: Context, Challenges, Cases*. Springer Singapore.

Agha, N. (2014). Success Factors of Startup Companies. An Empirical Analysis of E-Business Startups in North America. Retrieved July 29, 2022, from GRIN: https://www.grin.com/document/355234

Aggarwal, C. C. (2015). Data Mining: The Textbook. New York: Springer International Publishing.

Aktan, S. (2011). Application of machine learning algorithms for business. Investment Management and Financial Innovations, 8(2), 52-65.

Alpaydin, E. (2014). Introduction to Machine Learning. Cambridge: MIT Press.

Alvin, C., Peterson, B., & Mukhopadhyay, S. (2019). Static generation of UML sequence diagrams. *International Journal on Software Tools for Technology Transfer*, *23*(1), 31–53. https://doi.org/10.1007/s10009-019-00545-z

Amne. (2021, September 2). *Problems Facing Startups in Africa*. Shikana Group. https://shikanagroup.com/2021/09/02/problems-facing-startups-in-africa/

Bergström, G., Hujainah, F., Ho-Quang, T., Jolak, R., Rukmono, S. A., Nurwidyantoro, A., & Chaudron, M. R. V. (2022). Evaluating the layout quality of UML class diagrams using machine learning. *Journal of Systems and Software*, *192*, 111413. https://doi.org/10.1016/j.jss.2022.111413

Blank, S. G., & Dorf, B. (2020). *The start-up owner's manual: the step-by-step guide for building a great company. Vol. 1*. John Wiley & Sons, Inc.

Biau, G., & Scornet, E. (2015). A Random Forest Guided Tour. Test, 25(2), 1-43.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Berlin, Germany: Springer Science.

Blank, S. (2010). What's A Start-up? First Principles. Retrieved from Steve blank: https://www.grin.com/document/355234

Butler, J.E., Doktor, R., & Lins, F. (2010). Linking international entrepreneurship to uncertainty, opportunity discovery, and cognition. International Journal of Entrepreneurship, 8, 121–134.

Carrasco, O. C. (2019, April 04). Support Vector Machines for Classification. Retrieved from https://towardsdatascience.com/support-vector-machines-for-classification-fc7c1565e3

Carstoiu, D., & Grigorescu, C. (1995). OOAD Methods for O-O and KBS Development. *IFAC Proceedings Volumes*, *28*(24), 263–268. https://doi.org/10.1016/s1474-6670(17)46560-8

Chernev, B. (2022, April 28). What Percentage of Start-ups Fail? [30+ Stats for 2022]. Retrieved from https://review42.com/resources/what-percentage-of-start-ups-fail/

Conger, S. A., & Mason, R. O. (2013). Systems Analysis. *Encyclopedia of Operations Research and Management Science*, 1523–1536. https://doi.org/10.1007/978-1-4419-1153-7_1032

da Silva, F. R., & Bento, R. (2018). Predicting start-up success with machine learning". Lisbon, Portugal: Unpublished MS thesis, Dept. Inf. Manage., Universidade Nova do Lisboa.

Danilov, A. (2016). Initial Public Offering: The EU Prospectus Regime. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2844465

Daubie, M. & Meskens, N. (2012). Business Failure Prediction: A Review and Analysis of the Literature. In C. Zopounidis, New Trends in Banking Management. Contributions to Management Science. Physica, Heidelberg.

de Lange, P., Nicolaescu, P., Neumann, A. T., & Klamma, R. (2020). Integrating Web-Based Collaborative Live Editing and Wireframing into a Model-Driven Web Engineering Process. *Data Science and Engineering*, *5*(3), 240–260. https://doi.org/10.1007/s41019-020-00131-3

Dellermann, D., Lipusch, N., Ebel, P., & Michael, K. (2017). Finding the unicorn: Predicting early-stage start-up success through a hybrid intelligence method. Thirty-Eighth International Conference on Information Systems, South Korea 2017, (pp. 1-12).

Depren, S.K., Aşkın, Q. E, & Öz, E. (2017). Identifying the Classification Performances of Educational Data Mining Methods A Case Study for TIMSS. Journal of Educational Sciences: Theory & Practice, 17(5), 1605–1623, DOI https://doi.org/10.12738/estp.2017.5.0634.

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. BMC Medical Genomics, 4(1). https://doi.org/10.1186/1755-8794-4-31

Douglas, J., Douglas, A. & Muturi, D. (2017). An Exploratory Study of Critical Success Factors for SMEs in Kenya. The 2nd International Conference University of Verona, Verona (Italy) (pp. 1-13). Verona: University of Verona.

Faria, J. (2022). A number of tech start-ups that raised funding in Kenya 2015-2021. Retrieved from https://www.statista.com/statistics/1279467/number-of-funded-start-ups-in-kenya/

Freund, Y, & Mason, L. (1999). The Alternating Decision Tree Algorithm. Proceedings of the 16th International Conference on Machine Learning, (pp. 124-133).

Frost, J. (2022, July 7). *Cronbach's Alpha: Definition, Calculations & Example*. Statistics by Jim. https://statisticsbyjim.com/basics/cronbachs-alpha/

Grummer, J.-M. (2013). Was Startups wert sind. Retrieved from Gruenderszene : http://www.gruenderszene.de/allgemein/unternehmensbewertung-startups-teil-1 [Accessed July 26, 2013].

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. (1998). Support vector machines. 13(4),. IEEE Intelligent Systems and their applications, 13(4), 18-28.

Hedrick, T.E., Bickman, L., & Rog, D. (2013). Applied Research Design: A Practical Guide. London: Sage Publications, inc.

Huang, B. (2016). Predict Start-up Success using Network Analysis and Machine Learning Techniques. Stanford: Stanford University.

IIango, G & Kumar, S. (2017). Flower Species Recognition System using Convolution Neural Networks and Transfer Learning. . 4th International Conference on Signal Processing, Communications and Networking (pp. 1-6). Chennai: ICSCN.

Jackson, T. (2022). The African Tech Start-ups Funding Report: African tech start-up funding in 2021 – more and more for the "big four, (2021). Disrupt Africa. Retrieved from https://disrupt-africa.com/2022/02/04/african-tech-start-up-funding-in-2021-more-and-more-for-the-big-four/

Kidder, D. S. (2013). The Start-up playbook: Secrets of the fastest-growing start-ups from their founding entrepreneurs. Chronicle Books. Kleinbaum, D. G. and Klein, M. (2012). Logistic Regression. A Self-Learning Text (2rd ed.). London: Springer.

Kollmann, T., & Kuckertz, A. (2003). E-Venture-Capital. Gabler Verlag.

Korting, T. (2014). C4. 5 algorithm and Multivariate Decision Trees. Brazil: Image Processing National Institute for Space Research – INPE.

Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the Outcome of Start-ups: Less Failure, More Success. IEEE 16th International Conference on Data Mining Workshops (pp. 798–805.). IEEE Computer Society.

Lacave, C., & Diez, F. (2012). A review of explanation methods for Bayesian networks. The Knowledge Engineering Review. Knowledge Engineering Review, 17, 107–127.

Leung, K. (2016). Naive Bayesian classifier. Retrieved from http://www.sharepdf.com/81fb247fa7c54680a94dc0f3a253fd85/naiveBayesianClassifier.pdf,

Li, J. (2020). Prediction of the Success of Startup Companies Based on. Tianjin: Li.

Liaw, A., & Wiener, M. (2012). Classification and regression by randomForest. R News, 2(3), 18-22.

MacMillan, I. C., Siegel, R., & Narasimha, P. S. (1985). Criteria used by venture capitalists to evaluate new venture proposals. *Journal of Business venturing*, 1(1), 119-128.

Makarenko, E.N., Chernysheva, Yu., G., Polyakova, I.A. & Makarenko, T. (2019). The Success Factors of Small Business. International Journal of Economics and Business Administration, 7(2), 280-288.

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I. & Mendonca, A. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity, and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees, and random forests. BMC, 4(299), 1-14.

Mishra, S.B., & Alok, S. (2011). Handbook of Research Methodology: A Compendium for Scholars & Researchers. New Delhi: Educreation Publishing.

Mitchell, T. M. (1997). Machine learning (1st ed.). USA: McGraw-Hill, Inc.

Moniruzzaman, A. B. M., & Hossain, D. S. A. (2013). Comparative Study on Agile software development methodologies. arXiv preprint arXiv:1307.3356.

Mugenda, O. M., & Mugenda, A. G. (2003). Research methods, qualitative and quantitative approaches. Nairobi: African Centre for Technology Studies.

Mule, Dr. S. S., & Waykar, Y. (2015, January). *(PDF) role of use case diagram in software development*. ResearchGate.

https://www.researchgate.net/publication/322991847_role_of_use_case_diagram_in_software_development

Murimi, B. (2014). Factors Affecting the Success of Start-Up Of Youth Enterprises In Nairobi County, Kenya. Nairobi, Kenya: Unpublished Thesis, University of Nairobi.

Narasimha Murty, M. & Susheela Devi, V. (2011). Pattern Recognition: An Algorithmic Approach. New York: Springer Science & Business Media.

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., Malík, P., & Hluchý, L. (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, *52*(1), 77–124. https://doi.org/10.1007/s10462-018-09679-z

Njuguna, J. (2019). *Factors Inhibiting The Success of Entrepreneurial Intentions on Launching a Business: A Case of United States International University -Africa Students*. https://philarchive.org/archive/njufit

Okrah, J., Nepp, A. & Agbozo, E. (2018). Exploring the factors of start-up success and growth. The Business and Management Review, 9(3), 229-337.

Oppermann, A. (2021, February 28). What is Deep Learning and how does it work? | Towards Data Science. Medium. https://towardsdatascience.com/what-is-deep-learning-and-how-does-it-work-2ce44bb692ac

Park, kyoung J. (2017). A Study on Improvement of BIM(Building Information Modeling) Working Environment in Architectural Design Area Using API(Application Program Interface). *KOREA SCIENCE & ART FORUM*, *30*, 107–117. https://doi.org/10.17548/ksaf.2017.09.30.107

Picken, J. C. (2017). From start-up to scalable enterprise: Laying the foundation. Business Horizons, 60(5), 587–595. http://dx.doi.org/10.1016/j.bushor.2017.05.

Piryonesi, S., Madeh, El-Diraby, & Tamer, E. (2020). Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. Journal of Transportation Engineering, 146(2), 04020022. doi:10.1061/JPEODX.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury Google, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., Devito, Z., Raison Nabla, M., Tejani, A., Chilamkurthy, S., Ai, Q., Steiner, B., & Facebook, L. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. https://arxiv.org/pdf/1912.01703.pdf

Ratner, B. (2017). Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data. Florida, United States: CRC Press.

Ries, E. (2011). Lean start-up: Today's entrepreneurs use continuous innovation to create radically successful businesses. New York: Crown Business.

Rogers, S. (2020). Entrepreneurial Finance, Fourth Edition : Finance and Business Strategies for the Serious Entrepreneur, 4th Edition. McGraw-Hill.

Rouhani, S., & Lecic, D. M. (2018). *Business Intelligence Impacts on Design of Enterprise Systems*. Encyclopedia of Information Science and Technology, Fourth Edition. https://www.igi-global.com/chapter/business-intelligence-impacts-on-design-of-enterprise-systems/184005

Saleh, M. (2022). Africa: start-up failure rate by country. Statista. https://www.statista.com/statistics/1295678/startup-failure-rate-in-africa-by-country/

Shepherd, D. A. (1999). Venture capitalists' assessment of new venture survival. Management science, 45(5), 621-632

Skawi´nska E., & Zalewsk, R. (2020). Success Factors of Start-ups in the EU—A Comparative Study. Sustainability, 12(8800), 1-28, doi:10.3390/su12198200.

Somekh, B., & Lewin, C. (Eds.). (2005). *Research methods in the social sciences.* Sage.

Statista. (2020). Africa: number of start-ups by country. Statista. https://www.statista.com/statistics/1290679/number-of-startups-in-africa-by-country/

Swartout, W. (1986). Knowledge Needed for Expert System Explanation. Future Computing Systems, 99–114.

Taherdoost, H. (2016). (PDF) Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. ResearchGate. https://www.researchgate.net/publication/319998004_Validity_and_Reliability_of_the_Research_Instrument_How_to_Test_the_Validation_of_a_QuestionnaireSurvey_in_a_Research

Tandon. (2010). Research methodology: methods and techniques. Anmol Publications Pvt Lt.

Tao, Y., & Kung, C. (1991). Formal definition and verification of data flow diagrams. *Journal of Systems and Software*, *16*(1), 29–36. https://doi.org/10.1016/0164-1212(91)90029-6

Tomy, S., & Pardede, E. (2018). From Uncertainties to Successful Start-ups: A Data Analytic Approach to Predict Success in Technological Entrepreneurship. Sustainability, 10(602), 1-24, doi:10.3390/su10030602.

Unal , C., & Ceasu, I. (2019). A Machine Learning Approach Towards Startup Success. Berlin: n.d.

Verikas, A., Vaiciukynas, E., Gelzinis, A., Parker, J. & Olsson, M. (2016). . Electromyographic Patterns during Golf Swing: Activation Sequence Profiling and XC Prediction of Shot Effectiveness. 2016 April 23;16(4):. Sensors (Basel)., 16(4), 592. DOI: 10.3390/s16040592.

Wakiaga, P. (2020, April 28). SMEs are critical in attaining the manufacturing dream. Retrieved from Kenya Association of Manufacturers: https://kam.co.ke/smes-critical-in-attaining-manufacturing-dream/

Watson, K. & Hogarth-Scott, K. (1WE998). Small business start-ups: success factors and support implications. International Journal of Entrepreneurial Behaviour & Research, 4(3), pp. 217-238.

Weil, D. (2021, April 29). The Risks and Rewards of Angel Investing. The Wall Street Journal. Retrieved from https://www.wsj.com/articles/risk-rewards-of-angel-investing-11635957307

Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. (2017). Data Mining: Practical Machine Learning Tools and Techniques (4th ed.). London: Elsevier.

Wooldridge, J. (2015). Introductory Econometrics: A Modern Approach (6th ed.). United Kingdom: Cengage Learning.

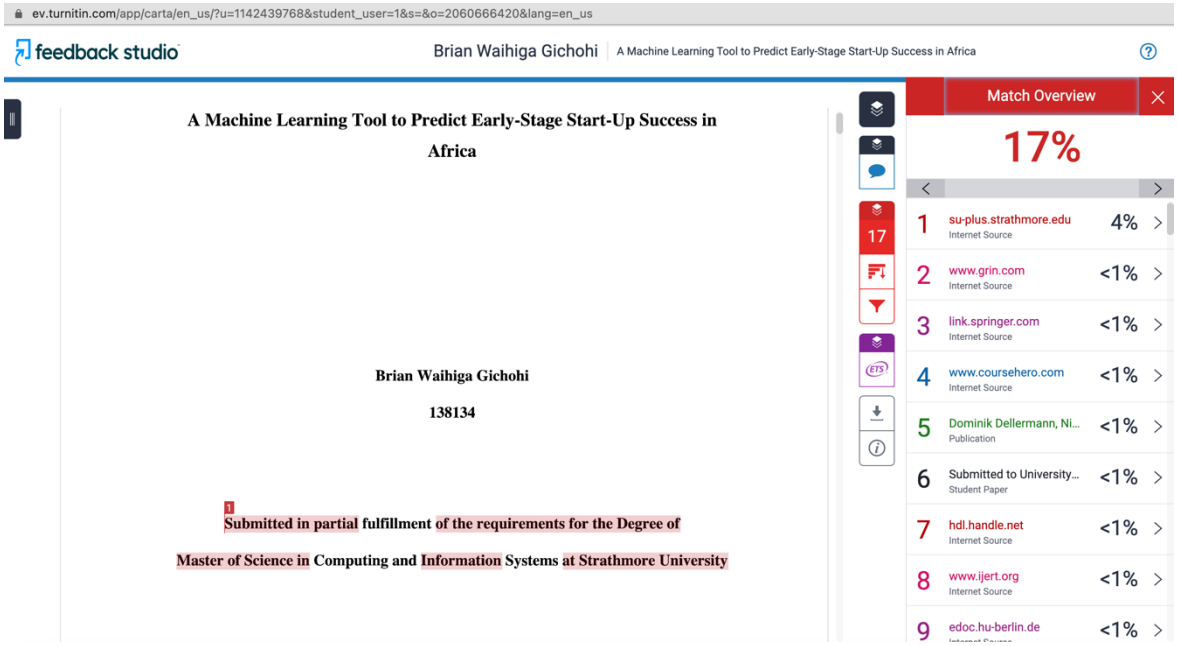
Yu, L., Li, B., & Jiao, B. (2019). Research and Implementation of CNN Based on TensorFlow. *IOP Conference Series: Materials Science and Engineering*, *490*, 042022. https://doi.org/10.1088/1757-899x/490/4/042022

Zacharakis, A. L., & Meyer, G. D. (2000). The potential of actuarial decision models: can they improve the venture capital investment decision? *Journal of Business venturing*, 15(4), 323-346.

Żbikowski, K. & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing and Management*, 58, 1-18

Appendices

Appendix I: Originality Report



Appendix II: Ethical Review

Strathmore UNIVERSITY

22nd February 2023

Mr Gichohi Brian Waihiga,
brian.waihiga@strathmore.edu

Dear Mr Gichohi,

RE: A Machine Learning Tool to Predict Early-Stage Start-Up Success in Africa

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU- master's** research proposal. Your application reference number is **SU-ISERC1584/23.** The approval period is from **22nd February 2023 to 21st February 2024.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, and 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 48 hours of notification
iv. Any changes, anticipated or otherwise, that may increase the risks or affect the safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 48 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.

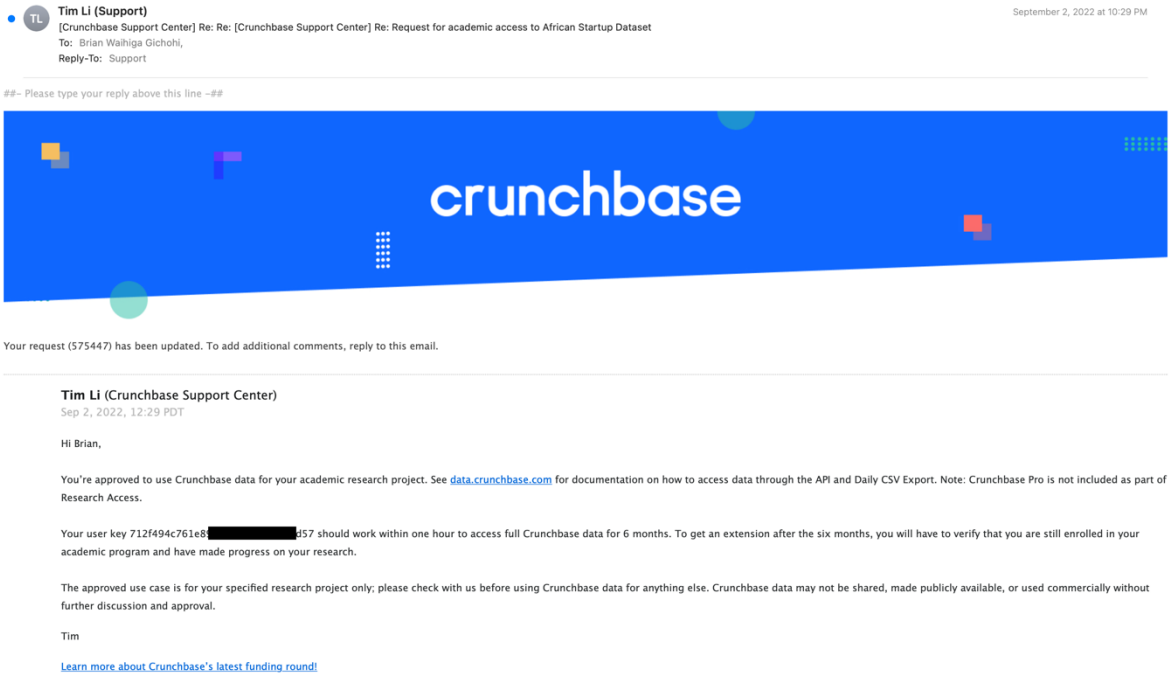
Yours sincerely,

for: **Dr Ben Ngoye,**
Secretary; SU-ISERC

Cc: Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
22-Feb-2023
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Appendix III: Data Use Approval from Crunchbase



Tim Li (Support) September 2, 2022 at 10:29 PM
[Crunchbase Support Center] Re: Re: [Crunchbase Support Center] Re: Request for academic access to African Startup Dataset
To: Brian Waihiga Gichohi,
Reply-To: Support

##- Please type your reply above this line -##

crunchbase

Your request (575447) has been updated. To add additional comments, reply to this email.

Tim Li (Crunchbase Support Center)
Sep 2, 2022, 12:29 PDT

Hi Brian,

You're approved to use Crunchbase data for your academic research project. See data.crunchbase.com for documentation on how to access data through the API and Daily CSV Export. Note: Crunchbase Pro is not included as part of Research Access.

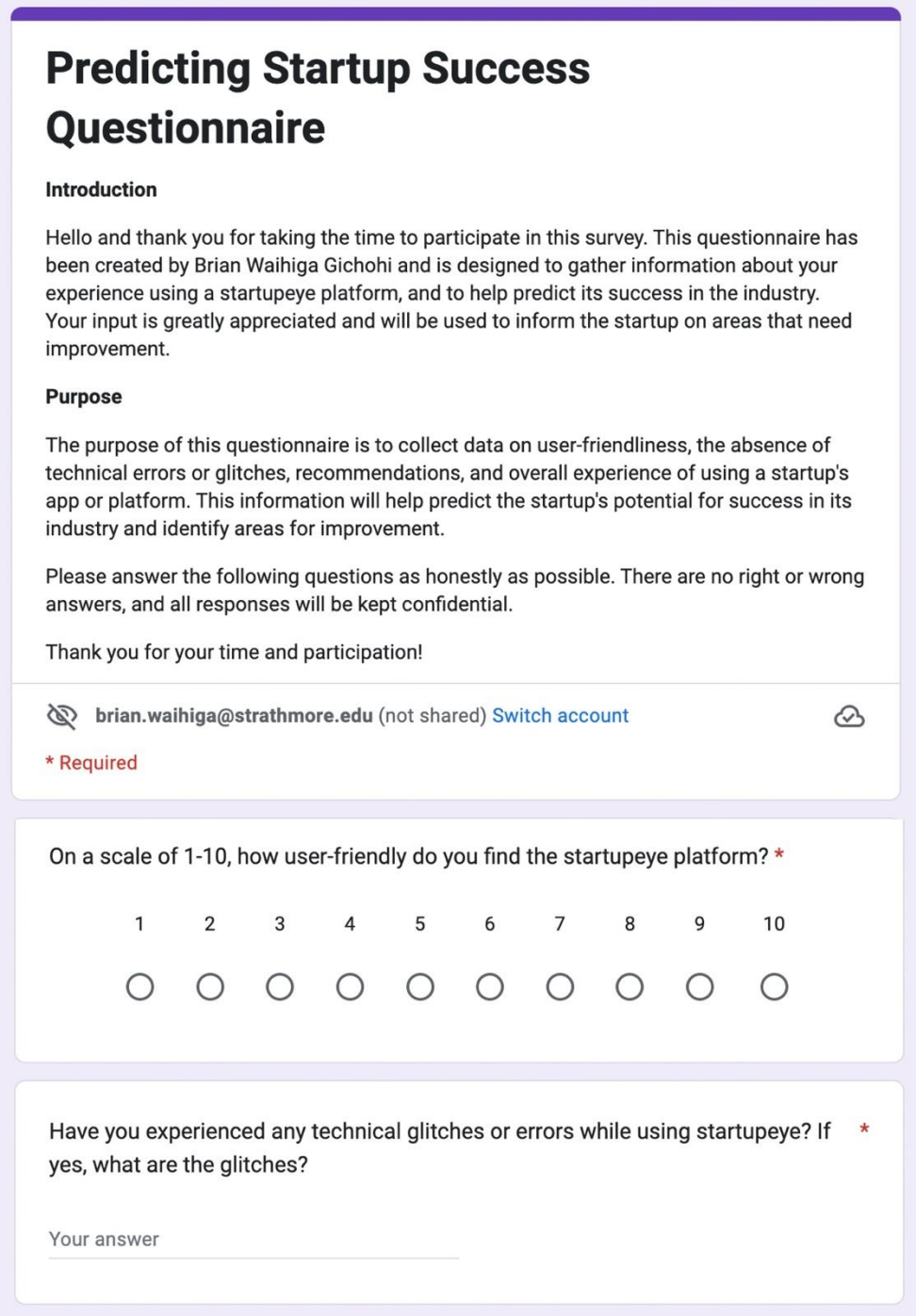
Your user key 712f494c761e8[redacted]d57 should work within one hour to access full Crunchbase data for 6 months. To get an extension after the six months, you will have to verify that you are still enrolled in your academic program and have made progress on your research.

The approved use case is for your specified research project only; please check with us before using Crunchbase data for anything else. Crunchbase data may not be shared, made publicly available, or used commercially without further discussion and approval.

Tim

Learn more about Crunchbase's latest funding round!

Appendix IV: Usability Testing Questionnaire



Predicting Startup Success Questionnaire

Introduction

Hello and thank you for taking the time to participate in this survey. This questionnaire has been created by Brian Waihiga Gichohi and is designed to gather information about your experience using a startupeye platform, and to help predict its success in the industry. Your input is greatly appreciated and will be used to inform the startup on areas that need improvement.

Purpose

The purpose of this questionnaire is to collect data on user-friendliness, the absence of technical errors or glitches, recommendations, and overall experience of using a startup's app or platform. This information will help predict the startup's potential for success in its industry and identify areas for improvement.

Please answer the following questions as honestly as possible. There are no right or wrong answers, and all responses will be kept confidential.

Thank you for your time and participation!

brian.waihiga@strathmore.edu (not shared) Switch account

* Required

On a scale of 1-10, how user-friendly do you find the startupeye platform? *

1 2 3 4 5 6 7 8 9 10

Have you experienced any technical glitches or errors while using startupeye? If yes, what are the glitches? *

Your answer

Would you recommend startupeye to your family and friends? If no, why not? *

Your answer

How satisfied are you with your overall experience using startupeye platform? *

	1	2	3	4	5	
Not Satisfied	○	○	○	○	○	Very Satified

Do you think startupeye platform is better than its competitors in terms of user-friendliness and lack of technical issues? If yes, why? *

Your answer

What suggestions do you have for the startup to improve user-friendliness and overall experience on its app or platform?

Your answer

Do you think the startupeye platform has the potential to become successful in its industry? Why or why not?

Your answer

Submit Clear form