ISSN: 2995- 4010

Research Article

Current Trends in Business Management

# Artificial Intelligence For Startup Risk And Investment Readiness Assessment: A Machine Learning Model From the African Innovation Ecosystem

**Abiodun Ajanaku***

*Department of Applied Artificial Intelligence and Data Science, Solent University, Southampton United Kingdom*

***Corresponding Author**
Abiodun Ajanaku, Department of Applied Artificial Intelligence and Data Science, Solent University, Southampton United Kingdom.

**Submitted:** 2025, Jun 06; **Accepted:** 2025, Jun 27; **Published:** 2025, Jul 08

**Citation:** Ajanaku, A. (2025). Artificial Intelligence For Startup Risk And Investment Readiness Assessment: A Machine Learning Model From the African Innovation Ecosystem. *Curr Trends Business Mgmt, 3*(2), 01- 21.

**Abstract**

*This study presents a novel approach to assessing startup risk and investment readiness in Africa using artificial intelligence and machine learning techniques. Despite the rapid growth of technology ecosystem across the continent, access to early-stage funding remains one of the most critical challenges for startups, major contributors to these challenges include but not limited to high perceived risk, fragmented information, investors uncertainty, lack of standardized, transparent, and scalable tools for assessing the viability and investment readiness of early-stage ventures. This research aimed to provide practical solutions to these challenges through the development and validation of a machine learning (ML) model tailored to assess startup risk and investment readiness within the African innovation ecosystem. The research utilized synthesized and anonymised multi-dimensional dataset of 10,000 startups modeled after small-scale real-world cases; constructed through domain expertise and rule-based logic that reflect industry benchmarks. It also integrated historical insights from different startup capacity building and business management programmes spanning 5 years. The dataset used in the study categorized startup risk across five dimensions: financial, operational, compliance, technology and strategic – reflecting the comprehensive due diligence process that investors typically conducted on ventures. A mixed method was deployed to include both qualitative and quantitative variables across the above stated five domains, starting with an initial 27 features, and narrowed down to 19 most influential features selected through statistical and data science techniques including Random Forest classifier feature selection, mutual information, correlation heatmap, and chi-squared test of association between the different independent variables and the target. The performance of five machine learning models was benchmarked including Random Forest (RF) classifier, HistGradient Boosting (HGB), K-Nearest Neighbors (KNN), Logistic Regression (LogReg) and Decision Tree Classifier (DT). Amongst the various ML models evaluated, Random Forest demonstrated the strongest predictive capability in multi- class risk classification tasks, with precision, recall, and F1 score all consistently averaged at 0.98 across classes of 1,500-test set. To minimize missed detection of high-risk startups (Type II errors), the model was optimised for high recall in the high-risk class ensuring a conservative and risk-averse screening strategy. The model categorized startup into three distinct risk clusters (Low, Medium, High) powered by unsupervised ML techniques K-Means and Elbow method. Cross validation was performed to assess the model performance and generalisation; hyperparameter tuning was conducted using Random Search. The results and findings of the research validated that artificial intelligence cannot only predict risk levels and investment readiness with high accuracy, but also uncover nuanced drivers of startup viability in underserved markets such as the African innovation ecosystem. The model prioritizes safety by ensuring that no high-risk startups are misclassified as investable, effectively eliminating false negatives (Type II errors). It also reduces the likelihood of classifying a promising, investable startup as high risk - false positives (Type 1 errors). This conservative approach reduces exposure to poor investment decisions, making it ideal for use in early-stage screening or high-stakes risk assessment. Furthermore, AI/ML models can drive more inclusive and data-driven investment decisions using multi-dimensional business information. Ultimately, this research contributes to the growing body of knowledge in the field*

*of applied AI in finance and enterprise support programmes – providing practical tools to investors, accelerators, founders, policy makers and other stakeholders to de-risk innovative enterprises, provide technical assistance and accelerate effective resource allocations.*

**Keywords:** Artificial Intelligence, Machine Learning, Investment Readiness, Business Risk Prediction, Technology Startups, African Innovation Ecosystem, Startup Financing, AI in Emerging Markets, Predictive Modeling

## 1. Introduction and Background

In Africa's rapidly evolving innovation ecosystem, thousands of startups face challenges raising funding and scaling sustainably dues to several reasons including but not limited to high perceived business risks, unclear and unreliably financial projections, high burn rate without a clear growth strategy, lack of insights into what is driving performance and weak financial governance. Technology ("tech") start- ups as well as any traditional business face these challenges which can significantly impact their survival and growth if not detected and mitigated early. Furthermore, a large percentage of startups are inherently associated with high-failure rate and one of the reasons why tech startups fail beyond funding is absence of a comprehensive due diligence to establish if the company is investment worthy. Around 20% of startups fail within the first two years. At the five-year mark, almost half (49.7%) have shut down and out of business. After ten years, more than 65% have closed their doors. Only 25% of businesses make it beyond 15 years [1]. However, the few high-performing ones have the potential to ultimately generate high reward for investors. Hence, every investor before signing an investment agreement wants to gauge and understand the business advantages, risks, and drawbacks.

### 1.1 Definition of Key Concepts

#### 1.1.1 Startups

Startups are widely recognized as catalysts for innovation, job creation and economic growth across both developed and emerging markets. Startups are typically defined by as early-stage businesses, innovation-driven, associated with high uncertainty and experimentation; they are largely in pursuits of scalable and repeatable business models. While startups operate in diverse industry, this study focuses on technology-enabled startups - product or service focus companies that are developing or have developed innovative products, services, or solutions based on new or emerging technologies to solve pressing socio-economic challenges. They can fall into any one of these funding or business categories: Idea Stage (Pre-Seed Stage), Seed Stage, Early Stage (Series A), Growth Stage (Series B), Expansion Stage (Series C, D and Beyond), Maturity Stage / Exit. The goal of most founders of tech startups is to attract the right investors and land the required investment to drive operations, achieve a product-market fit and scale the business model. Given the peculiar characteristics and mode of operations of most tech startups, 99% of tech startups require funding at different stages of their business lifecycle; ranging from early, growth, expansion, maturity, and up to exit stage [2]. Depending on the stage of development, funding can come from friends and family or take the form of pre-seed, seed amongst others during the start-up stage. On the other hand, funding can take the form series A to D and beyond during the scale up stages of the funding journey and finally exiting through public offerings or selling the company. These funding are typically required for research and development, development of minimum viable product (MVP), finding product-market fit, further development and improvement of product, marketing/building customer base, scaling operations, expansion into a new market and preparing for Initial public offer (IPO) or acquisition.

#### 1.1.2 Investment Readiness

Investment readiness is broadly defined as having data-driven clarity relating to a business model, financial projections, business risks, financial transparency, legal/compliance, clear market, and distribution strategy. In other words, it refers to the extent a startup demonstrates the capacity and preparedness to attract, secure, and effectively deploy external funding. In this study, investment readiness is operationalized as a classification outcome embedded into a risk score derived from quantitative indicators that aligns with investors' expectations in early-stage African startups [2]. This study simulates aspects of due diligence using data-driven variables to enhance the efficiency and objectivity of startups evaluation.

#### 1.1.3 Business Risks

A business risk is any controllable or uncontrollable event or condition which occurrence could have a negative impact on the business outcomes, objectives, and key results. In other words, they are multi- dimensional outcomes that could be predicted through historical and real-time data, serving as proxy for gauging viability and enhancing investors' confidence. Controllable risks are possible risks that can be influenced, managed, or mitigated through a proactive measure. Business risks included in this category includes but not limited to financial, operational, strategic, management, technology, compliance, customer risks. These risks could be managed by putting in place the necessary mitigation plan and processes [3]. On the other hand, the uncontrollable risks are driven by external factors beyond the control of a business. This risk category is unpredictable; when they occur, businesses need to develop strategies to adapt or mitigate their impact rather than prevent them completely. Uncontrollable risks comprise of economic and market risks such as inflation/deflation, interest rate, foreign currency fluctuations; political and regulatory including policy changes, political instability; natural and environment risks such as the COVID-19 pandemic, natural disasters and climate change; technology risk resulting from change in global trends and technology advancement; social risks resulting from change in customer preferences, demographics; global and geopolitical risks such as wars, conflicts amongst others.

### 1.1.4 Risk Management

In the context of this study, risk management is defined as a systematic process of identifying, evaluating and mitigating uncertainties that could negatively impact a business objectives or performance. Overall, controllable business risks could be managed proactively, if done effectively it could help startups build resilient, improve efficiency, position the business for long-term completive advantage and success. Controllable business can be managed using four (4) major approaches: Identification of risks through internal analysis of operations, processes, and systems. Secondly, once a risk is identified, the next step is to assess or evaluate the risk by gauging it likelihood and the impact of the risks on the business. The likelihood and impact of the risks will establish their severity on business outcomes, objectives, and results. Furthermore, after evaluating the risks, a risk mitigation plan needs to be put in place and may include clear call to action such as setting up a clear policies and procedures as mitigants, investment in training and technology, and monitoring performance and compliance. This may also extend to putting in place an insurance policy as a way of transferring the risks to a third party to indemnify the business if the risks occur, hiring legal and compliance experts to navigate regulations. The final step to any risk management is to continuously review, monitor, and update the mitigation strategies as necessary [3].

### 1.1.5 Due Diligence

Due diligence (DD) is an extensive examination of the internal and external aspects of a business venture for the purpose of potential investment, acquisition, or merger. It is typically carried out by an investor, or a third-party professional consultancy firm. The objective of due diligence is to prove or disprove claims made by the founders or management of a venture by looking at hard evidence on key metrics such as unit economics, customer churn rate, operational efficiency, resilience of the company, financials (revenue, cost, assets, liabilities, equity), legal, compliance status, product commercialisation process and other key fundamentals of the company [4]. There is a strong relationship between the stage a business lifecycle and the duration of the due diligence. The due diligence for tech startups at the scale up stage (growth, expansion and mature) typically take longer because of the large volume of information to examine. Hence, an investor's focus may also shift depending on the stage of the round. On the other hand, for tech start-up at the early stage (pre-seed and seed), the DD process may take less time due to less data and limited legacy information relating to revenue, cost amongst others. Hence, there is heavy reliance on the founders and management team to support the DD process. The DD process is usually classified into two distinct stages: the first part of the DD examines the commercial and technology part of the company and usually takes place before a term sheet is executed; the in-depth examination of financial and legal review may happen after the execution of the term sheet [4].

### 1.1.6 Artificial Intelligence (AI)

Artificial Intelligence is the branch of computer science that focuses on building systems capable of replicating human intelligence – such as learning (data driven inference), reasoning (rule-based logic), perception, and decision-making. AI is self-learning in detecting patterns in big data by clustering information that can be deployed in identifying relevant features [5,6]. In this research, AI denotes the conceptualisation, development and deployment of intelligent systems that are capable of learning from complex and multi-dimensional data of business to evaluate investment readiness and risk more objectively and at scale.

### 1.1.7 Machine Learning (ML)

Machine Learning is a core branch of AI and extension of traditional statistical techniques and mathematics that uses computational resources to develop algorithms. ML enables systems to learn patterns from historical big data, improves predictive performance overtime, and make decisions or predictions without explicitly being programmed [7]. ML techniques include but not limited to supervised (classification, regression), unsupervised (clustering, and ensemble methods) and reinforcement learning. This research utilised supervised machine learning – especially classification algorithms to develop predictive model that assesses and classifies startup business risk based on labelled datasets into LOW, MEDIUM, or HIGH. In addition, it gauges investment readiness.

## 1.2 Problem Statement

Despite the growing importance of startups in driving innovation and economic development in Africa, the continent's innovation ecosystem is characterised and constrained by a lack of practical, data-driven, and context-aware framework for evaluating the startups financial health, risk, and investment readiness. Hence, many startups fail to scale or sustain operations, not necessarily due to lack of innovation, but because investors and support institutions struggle to access their viability using real- time, data-driven insights. Startups will be required to pass due diligence before an investor can back the venture with funding or be classified as an investable venture. The traditional investment assessment models, largely developed with limited context awareness – do not fully account for operational realities, market volatility, infrastructure gaps, and socio-economic nuances inherent in the African startup ecosystems. In addition, in terms of the current state-of-the-art, most DD relating to investment in tech startups are mostly carried out by the "Big Four" audit and investment advisory firms and other global professional services powerhouses. This process requires large volume of information that covers several aspects of a business such as commercial, technology, financial and legal information amongst others which are aggregated and stored in a virtual data room. The challenge with the current state of art is that the process is time-consuming, very expensive for the potential investors and depending on the stage; it could take between 3-6 months (or more) to complete the process from start to end. The process unofficially begins at the first conversation between the company and the investor – which can be many years before and officially kick-starts during the fundraising conversation and journey. Hence, this could result in increased costs for clients if not delivered within scope, timeline and standard required. The above lack of predictive models tailored to African tech ecosystem ultimately leads to misallocation of capital, increased investment

risks, and diminished support for high-potential ventures. These challenges have necessitated an urgent need for intelligent system that considers both the quantitative and qualitative information to assess startups and business viability with contextual accuracy, provides actionable insights to stakeholders, and foster an environment of evident-based decision-making.

### 1.3 Significance of the Study

The acceleration of technology startups across Africa represents a transformative force in the continent's innovation, digital and economic development landscape. However, many of these startups struggle to secure funding due to several factors including but not limited inherently high risk, limited operating history, unproven business models, and the absence of standardized data-driven risk assessment tools. Risk modelling approach has traditionally relied on qualitative assessments and financial ratios, but these methods often fail to fully capture real-time signals or non-financial unique context and potential of early-stage ventures operating in the African innovation ecosystem. This study is significant because its leverages the artificial intelligence (AI) and machine learning (ML) to develop a predictive risk intelligence for assessing business risk and investment diligence specifically for African technology startups building the next generation digital infrastructure. Furthermore, the study contributes to the growing body of research that drives the application of emerging technologies such as AI/ML, and advanced data science techniques to real-world investment decision-making and ultimately bridging the gaps between traditional venture evaluation and intelligent automation.

By contextualising the model within the Africa's innovation ecosystem, the insights from this study will provide a practical, localised, and scalable framework that can elevate investment, business, and economic outcomes for the below beneficiaries:

Investors (Angel, Venture Capital, Private Equity & Others): Improve transparency in startups evaluation for investment and enhance investors' confidence. The AI/ML predictive system which is one of the outcomes of the research will help investors investing in the tech ecosystem to identify risks associated with a venture early with a high level of precision. In addition, it will facilitate a data-driven decisions, ultimately improving returns and reducing exposure to unforeseen risks.

- **Technology startups:** Support startups founder identifying and mitigating risks early before it cost them funding. The predictive system will help technology startups become investment ready, navigate investment due diligence process by uncovering potential risks that can deter investors and recommend possible mitigants to reduce exposure. The customized risk profile report generated by the predictive system can be used by tech startups as a risk register/manual during fund raising journey.
- **The Africa digital technology ecosystem:** The predictive risk intelligence system will serve as a 360-degree view of the business risks associated with a tech venture and transform investment due diligence by enhancing speed, accuracy, and depth of analysis. This holistic analysis uncovers insights and patterns in a venture that traditional approach might overlook.
- **Government and Public Policy:** The predictive risk intelligence system will strengthen public policy formulation and program targeting for startups in the Africa innovation ecosystem.

### 1.4 Research Objectives

The overarching aim of this research is to design and validate a machine learning model that can predict, classify business risk levels, and ultimately gauge the investment readiness of technology startups within African innovation ecosystem.

Consistent with the aim, the research specifically seeks to achieve these objectives:

- **Review** related works on AI/ML applications in startup risk evaluation and investment readiness.
- **Identify, engineer, and establish relationship amongst relevant features** such as financials, operational, compliance and strategic contexts influencing startup risks and investment decisions in the African Innovation ecosystem.
- **Develop a predictive machine learning model** that classifies startups based on risk level which signals investment readiness.
- **Validate the model's accuracy, performance and interpretability** using relevant and industry-accepted evaluation metrics such as confusion matrix; mutual information, correlation matrix and ensemble methods for features selection techniques and explainability.
- **Compare the model performance with traditional evaluation methods**, highlighting the added value of AI in investment decision making.
- **Provides practical insights and recommendations** for founders, investors, and policy makers based on the model's output and trends observed.

## 2. Literature Review

Technology ecosystems across Africa have witnessed significant growth over the past few years, mainly boosted and driven by a combination of factor such as increased internet connectivity, a youthful population, a growing entrepreneurial ecosystem, and torrent of venture funds, development finance, corporate involvement, as well as ever-growing, innovative communities. African tech startups have seen a significant increase in venture capital funding. According to Data trackers cited in, the amount raised by African startups in 2023 was between $2.9 billion and $4.1 billion, from $4.6 billion to $6.5 billion in total funding the previous year; fintech led the pack, accounting for nearly half of the total funding. Nigeria, Kenya, Egypt and South Africa are the top 4 largest beneficiaries [8]. The accelerating increase in venture capital (VC) funding and investors looking for ways to gauge associated risks with tech startups before making any investment commitment have resulted into more research on the risks associated with tech startups in the past few years. The increasing demand for AI/ML in investment due diligence and VC needing a more scientific and efficient model to understand the risks associated with tech startups have facilitated this research. Many recent studies have focused more broadly on the applications of AI/ML from a qualitative research perspective:

looking at the benefits and limitations of using CHATGPT and other AI models as a due diligence tools without any emphasis on the relevant features or baseline information to accurately predict the risk associated with tech startups. Hence, this study provides the opportunity for further investigation. To frame the foundation for this research, a systematic search was conducted using the Social Science Research Network (SSRN) database from inception to May 26, 2025 using the search string: "*Artificial Intelligence*" *OR* "*AI*" AND "*machine learning*" *OR* "*ML*" *AND* " *investment due diligence*" OR "due diligence" *AND* "*technology startups*" *AND* "*business risks*". The initial search yielded 10,000 unique papers of which 9,900 articles were excluded because they did not meet the inclusion criteria, were duplicates or were related to other studies. After applying relevance criteria – including empirical focus, use of AI/ML methods and relevant to startups financing. Full-text reviews were conducted for 100 articles and a final set of 6 articles were included for in-depth review. In addition to these core articles, supplementary materials from industry reports, books, relevant literatures were included to broaden the contextual and theoretical grounding of the study, particularly for African innovation ecosystems where academic literature remain relatively sparse.

| Authors(s) | Year | Focus Area | Methodology | Key Findings | Identified Gaps |
|---|---|---|---|---|---|
| Aziz and Dowling [9] | 2024 | AI and Machine Learning for Risk Management. | A non-technical review, analysis, and empirical evidence (Qualitative). | Presented an optimistic view of AI and machine learning in risk management fields such as credit risk, market risk, operational risk, and compliance ("Regtech") with some limitations around suitable data management policies, transparency, and lack of suitable skillsets within the firms. | No predictive model. A non-technical review with empirical evidence. |
| Laseter, Frazer and Boatright [10] | 2022 | Artificial Intelligence in Business: Machines and Management. | Technical note: an overview of artificial intelligence (Qualitative). | Highlighted opportunities for collaboration between AI and employees. In addition, it considers | Limited to application of AI to business. Not ML-based. |
| | | | | approach for responsible AI implementation. | |
| Petersons [11] | 2024 | Artificial Intelligence (AI) and machine learning in digitalization. | Detailed review of the impacts of AI and ML in digitalisation in various sectors (Qualitative). | Highlighted the practical applications of AI/ML in diverse sectors such as finance, healthcare, manufacturing, transportation, and retail. In addition, call for responsible deployment and ethical governance of AI technologies. | No focus on startup, investment readiness or AI/ML models. |
| Abikoye [12] | 2021 | Machine learning models and AI for predicting financial crises: applicability and accuracy. | Theoretical Framework: It compares the accuracy and reliability of various AI/ML models based on historical data. | The results showed that while all the AI/ML models performed well, the Neural Networks especially the convolutional neural networks (CNNs) and recurrent neural networks (RNNs), displayed a high accuracy due to their ability to model complex relationships. | No focus on startups or investment contexts. |

| | | | | | |
|---|---|---|---|---|---|
| Krause [13] | 2023 | ChatGPT and Other AI Models as a Due Diligence Tool: Benefits and Limitations for Private Firm Investment Analysis | Theoretical Framework on the benefits, and challenges of ChatGPT and other AI models. | Highlighted the benefits of machine learning techniques for due diligence on private companies. The research emphasised the importance of sentiment analysis for extracting insights from unstructured data and constructing a database of structured and unstructured dataset for evaluating private firms. | Although focus on private firms which include startups. However, no predictive models were demonstrated. |
| Sanz-Prieto, De-la-fuente-Valentín and Ríos-Aguilar [14] | 2021 | Technical due diligence as a methodology for assessing risks in start-up ecosystems. | Mixed methods including qualitative and quantitative approach. The sample included 30 experts, to whom a survey applied, and 10 of them, an interview that was subjected to a process of triangulation of the information, which supported by documentary arches. | The results showed the need to identify technological risks (product, service and process); commercial risks regarding the scalability of the business; and financial, legal, fiscal and environmental risks as part of a comprehensive and integral procedure. | Limited application of AI to business. Not ML-based. The study was based on deductive logic and applying survey to 30 domain experts. |
| Malhotra [15] | 2018 | AI, Machine learning & deep learning risk management & controls | Theoretical framework | Emphasized the urgent need for risk management controls in getting the best out of AI, but also ensuring that the worst fears about AI do not really come true. | Although focus on risk management. However, no startups, investment contexts. No predictive ML models was demonstrated. |
| Yu & Wang [16] | 2021 | Risk Prediction for SMEs | SVM and Gradient Boosting | Demonstrated accurate risk scoring using transaction data. | SME-focused, not early-stage tech startups. |
| Agbo [17] | 2022 | Predicting Start-up status using Funding and Sentimental data | Mixed: Quantitative publicly available financial information from Crunchbase and qualitative tweet and profile data from Twitter were analysed using eight specific machine learning algorithms which were all classifiers. | The study emphasized the impact of social media presence/online legitimacy of a start-up, has proved important to the overall performance of the model by being among the most important features required in the final model showing the crucial role this type of data plays in predicting start-ups status. | Applied ML/AI models. Focused on startups and investment contexts. However, the study focused more on the social media/online presence of startups in predicting status and funding. |

**Table 1: Summary of Key Literature Reviewed on AI/ML for Startup Risk and Investment Readiness**

### 2.1 Traditional Approaches to Risk Management and Investment Readiness

Early works such as Aziz and Dowling laid the conceptual foundation; explored a non-technical and analytical approach on how AI and ML is transforming the risk management [9]. Their methodology was based on a broader review of the application of AI/ML, the benefits and concerns in the managing different risks ranging from credit risk, market risk, operational risk, and

to compliance risk ('RegTech'). In their study, they highlighted the key players using AI/ML to automate, manage risks and drive regulatory compliance. These companies include IBM; Nordic KYC Utility: anti-money laundering infrastructure; Zest Finance: AI/ML infrastructure for customer and SMEs leading; BBVA (the second largest bank in Spain); Knight Capital: stock trading platform using AI/ML techniques) amongst others. They concluded that the time-consuming and costly nature of risk management will diminish significantly through the applications of AI/ML. However, they reiterated the final major issue around transparency and ethics which AI-driven solutions need to further address [9]. Laseter, Frazer and Boatright in their technical paper "Artificial intelligence in business: Machine and Management" highlighted the different schools of thought around AI's potential and the possible roles of AI in business. They argued that AI and ML have provides a good opportunity for collaboration between employees and AI and considered one approach around responsible AI implementation. However, they concluded that despite the huge opportunity that AI provides as a transformational technology with massive potential for positive global impact, over reliance on AI/ML systems not only disrupt human activity but also could have negative impacts in certain circumstances [10].

Petersons carried a detailed review of the impacts of AI and ML in digitalisation in various sectors including finance, healthcare, manufacturing, retail, and transportation. The paper examined the challenges and ethical considerations associated with the widespread application of AI and ML in in digital transformation initiatives. The benefits of AI and ML includes enhancing productivity and efficiency by automating repetitive routine tasks, streamlining workflows, and optimal resources allocation. Whereas some of the concerns associated with Al and ML were data privacy and security concerns, the convoluted decision-making process of most of AI and ML models termed as "black box" approach, task automation leading to fear of job displacement, and the seemingly lack of transparency, accountability, interpretability, and the potential for unintended consequences or algorithmic errors with AI/ML systems. The author concluded that AI and ML technologies have emerged as powerful drivers of digital transformation in various sectors. However, it is important to ensure the responsible deployment and ethical governance of AI and ML technologies to maximize their benefits while mitigating potential risks and addressing societal concerns such as privacy, bias, and job displacement. This will be only possible with collaboration amongst various stakeholders: policy makers, technologists, researcher, scientists, and others [11].

Abikoye carried out research to evaluate the applications of AI/ML techniques in in predicting financial crises. The author examined the performance and accuracy of five (5) different AI/ML models in predicting financial crises using historical financial data. The models used included: Logistic Regression, Decision Trees and Random Forests, Support Vector Machines, Neural Networks, Gradient Boosting Machines. The results showed that while all the AI/ML models performed well, the Neural Networks especially the convolutional neural networks (CNNs) and recurrent neural networks (RNNs), displayed a high accuracy due to their ability to model complex relationships. Despite, the high accuracy rate, the Neural Networks require substantial computational resources and large datasets [12]. Krause examined the impact of CHATGPT and other AI models as a due diligence tool. The author carried out a qualitative study of the various AI models; their benefits and limitations in performing due diligence for private companies. In addition, regulation A and Regulation CF private companies were used as samples of the study. The findings of the research showed that general AI models, such as ChatGPT, face several challenges in performing investment due diligence on non-public firms due to the absence of standardized data, publicly available information presence of complex ownership structures and dynamic markets. Hence, the role of human experts will remain critical in performing investment due diligence. The author concluded that both human expertise and AI/ML models will lead to informed decision making and more effective outcomes [13]. Sanz-Prieto, De-la-fuente-Valentín and Ríos- Aguilar analysed and proposed a technical due diligence as a methodology to assess risks in Start- up ecosystems. The authors deployed a mixed of quantitative approach, and the qualitative approach, supported by a literature review with bibliographic arches. Thirty (30) experts were survey; 10 out of the 30 were interviewed and subjected to a process of triangulation of the information and further supported by documentary arches. The findings of the research showed that key relevant factors such as: technological risks (product, service, and process); commercial risks regarding the scalability of the business; and financial, legal, fiscal, and environmental risks as part of a comprehensive and integral procedure for assessing risks in start-up ecosystems [14]. Malhotra in his paper emphasized the need for strong controls with the use of AI systems for it to achieve its desired goals without causing harm during the learning process, avoid misinterpreting or resisting human control. The author further argued that risk management controls is most critical in not only getting the best performance of AI/ML, but also ensuring that the worst fears about the AI do not really materialize [15].

### 2.2 Machine Learning Applications to Risk Predictions

More recent studies, including those by Yu and Wang [16], focused on credit risk assessment for small and medium-sized enterprises (SMEs) within supply chain finance. They implemented Support Vector Machine (SVM) and Backpropagation (BP) neural network algorithms to evaluate financial indicators and predict default risks. Based on the relevant concepts of the supply chain management budget model, it explores the main factors influencing the financial impact of SMEs and the benefits of the supply chain budget in solving problems expenditure of SMEs, support vector machine is mainly based on solving the main credit risks of small and medium-sized enterprises, such as poor information transparency, low credit and various risk unknown factors. BP neural network is an algorithm that considers the components of supply chain financial financing. The study achieved high accuracy rates, suggesting that these machine learning models are effective tools for financial risk assessment. Nonetheless, the research centered on SMEs in a specific economic context, and the models may require adaptation to address the complexities of startup environments, particularly

in regions with less formal financial infrastructures [16]. Agbo developed a machine learning model aimed at predicting startup status using funding and sentimental data by leveraging financial information from CrunchBase, one of the largest public databases for start-ups, with profile and tweet data from Twitter, one of the top and largest social media databases, to thoroughly examine, analyse and predict start-up status [17]. A total of 2,613,123 tweets (the largest-scaled sentiment data ever recorded in the literature) from over 40,000 start-up Twitter profiles were analysed using eight specific machine learning algorithms which were all classifiers. The study demonstrated that a start-up status can be precisely predicted using a trained machine learning model with online legitimacy as a gauge of social acceptance based on tweets and profile data from Twitter and in combination with the startup funding data. However, the model's applicability was primarily confined to startups status using publicly available data which may not captured the nuances and unique attributes such as data scarcity, perceived high-risk and differing market dynamics peculiar to private firms in emerging markets as the African Innovation ecosystem [17].

### 2.3 Research Gap and Justifications

While some progress has been made in applying machine learning and artificial intelligence to risk assessment and investment decisions, most models overlooked the unique contexts of startups operating in the African innovation ecosystem. These startups are associated with high business risks, informal markets, unique investors-founders dynamics, unstructured business, and legacy data. Furthermore, existing models, either assess risk or readiness in isolation – when the identified risks are high, medium, or low; the models do not provide target recommendations to mitigate and addressed the business risk. This study addresses this gap by proposing an integrated ML/AL model tailored to the African innovation ecosystem. Hence, assessing both startup risk and investment readiness simultaneously and putting this in the context of funding stage and the business lifecycle.

## 3. Methodology

This study adopted a mixed methods research design - qualitative insights from supporting 50 startups in the last 5 years in different incubation and acceleration programmes across African innovation ecosystem. In addition, the study integrated quantitative modeling to develop an AI/ML-based predictive system for assessing risk and investment readiness among African technology startups. The methodology followed five key phases: ***systematic literature reviews, data generation through simulation of real-world startups business data combined with domain expert insights and inputs, feature engineering, predictive modeling, and model development*** – as systematically outlined in Figure 1.

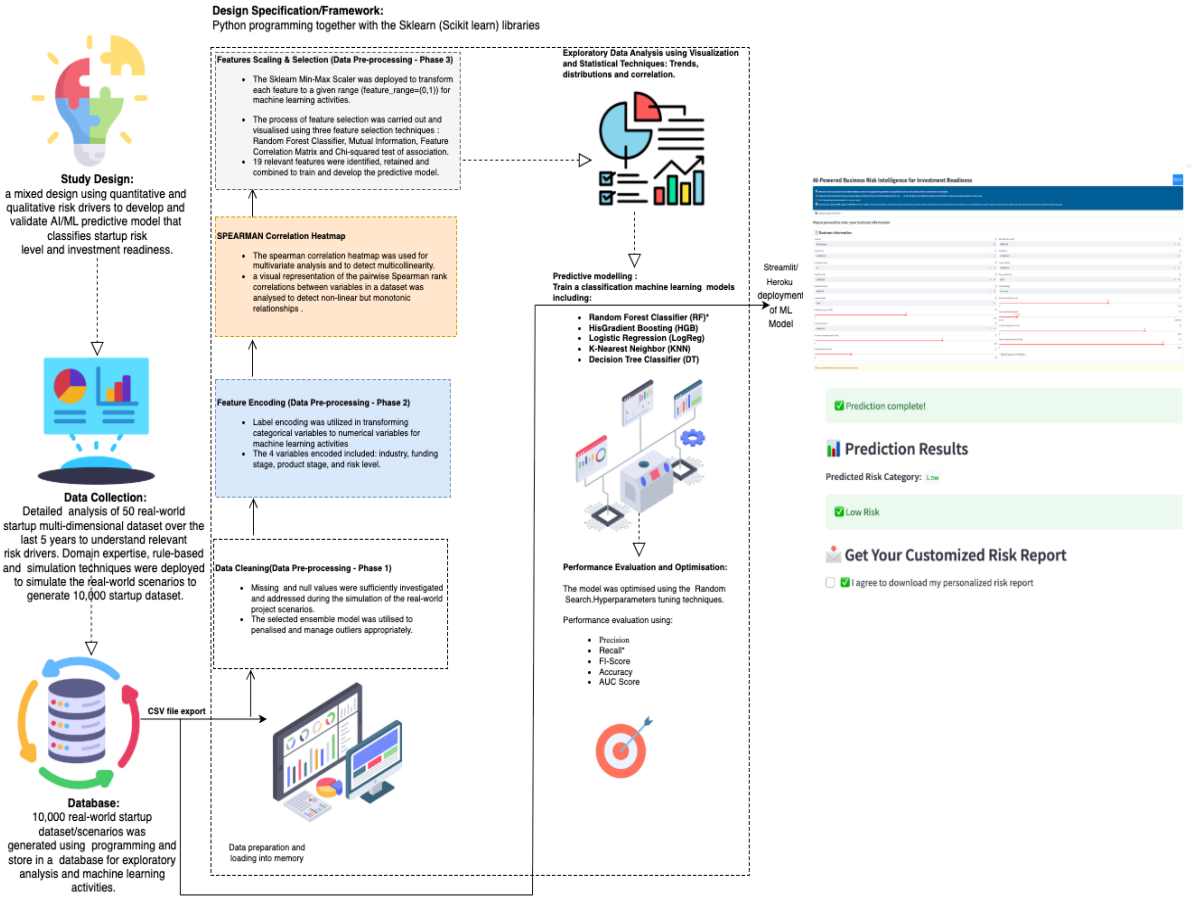


**Figure 1:** Schematic Representation of Research Methodology

### 3.1 Research Design and Data Collection

The research was based on systematic review of literature, domain experts' inputs, and reflects insights from literatures from startups databases, funding platforms, and regional reports from TechCruch amongst others. In addition to the insights from existing research, domain expertise was leveraged to define context-specific features and evaluate data quality. Where historical records were incomplete or inconsistent, data simulation techniques were applied to generate a synthetic dataset representing 10,000 African startups. Each instance (row) was assigned a unique identifier representing the business information for each startup. A total of 27 potential features were initially identified based on existing framework and experts-defined indicators of risks and investment readiness. These features were categorized into five dimension of startups risk aligning each risk type with a clear cause-effect relationships as detailed below:

- **Operational Risk:** In this study is measured by the likelihood and impact of inefficiencies or failures of internal processes. Represented by several variables including ***industry/sector, employee count, founders' strength score, funding stage, and team experience.***
- **Financial Risk:** This is measured by the likelihood and impact of cashflow instability or capital constraints. Represented by several quantitative variables including r***evenue, gross profit, gross margin, variable cost, fixed cost, customer acquisition cost (cac), burn rate, runway, current ratio, ebitda, ebitda margin, total asset, and total liabilities.***
- **Compliance/Regulatory Risk:** This is measured by the likelihood and impact of compliance challenges or policy shifts. Represented by qualitative feature such as ***risk level, intellectual property (IP) ownership, audit flags and data compliance.***
- **Strategic Risk:** This is measured by the likelihood and impact of misalignment between product and market. Represented by qualitative commercial metrics such as ***market traction, customer concentration, customer churn rate, and brand sentiment.***
- **Technology Risk:** In this study is measured by the likelihood and potential impact of failures or delays stemming from the ***stage of product development***, particularly where technology is untested, underdeveloped, or facing uncertainty in scalability and security. In other words, it includes the impact of failure in system architecture, functionality, and integration during the development lifecycle. The technology risk is measure by qualitative features product stage which includes idea, prototype, minimum viable product (MVP), Beta, launch, growth/scaleup and maturity/established. These risks are amplified during the early product development lifecycle where product design and technology stack are still being finalized or iterated.

### 3.2 Data Cleaning, Features Engineering, Exploratory Data Analysis (EDA) and Feature Selection

A comprehensive data preparation pipeline was implemented to ensure quality data input for machine learning activities and modeling:

- **Data Cleaning and Pre-processing:** The raw dataset underwent a comprehensive data cleaning and pre-processing to ensure quality data and model reliability. To address missing values within the dataset, a median imputation strategy was implemented for continuous and ordinal features. The median, rather than mean, was chosen due to its robustness against outliers, particularly in skewed distribution common amongst startup-related datasets. This approach ensured preservation of measures of central tendency of each variable without introducing bias from extreme values.

- **Encoding Categorical Variables:** The categorical variable risk score, representing qualitative assessment of startup risk levels (Low, Medium, High), was transferred using ordinal encoding (also known as label encoding) to preserve the inherent order of the categories. This shows an ordered relationship: Low < Medium < High. Each level was mapped to a corresponding to a corresponding integer (Low=1, Medium=2, High=3) to enable the machine learning model to interpret and learn from the rank-based relationship amongst risk categories. Similar techniques were applied to funding stage ("Pre-seed": 0, "Seed": 1, "Series A": 2, "Series B": 3, "Series C": 4, "IPO": 5) and product stage ("Idea": 0, "Prototype": 1, "MVP": 2, "Beta": 3, "Launched": 4, "Growth": 5, "Established": 6).

- **Creating Derived Variables:** Domain expertise was leveraged to introduce several derived variables into the dataset to reflect the unique attributes and nuances common with startup business dynamics such as "burn rate" (fixed cost + variable cost), "ebitda" (revenue – burn rate), "ebitda margin" (ebitda / revenue) * 100), "gross margin" ((revenue – variable cost) / revenue) * 100), "gross profit" (revenue – variable cost), "runway months" (cash reserve / burn rate)

- **Application of Domain Expertise to Incorporate Rule-Based Features for Startup Dynamics:** In addition to statistical pre-processing and encoding techniques, a rule-based feature engineering approach were deployed to capture nuanced dynamics of startup development. Domain expertise informed the creation of conditional rule that reflects typical patterns in startup funding stages, product maturity, ultimately associated risk profiles of each startup. For sample, startup at the pre-seed or seed stage with no or little revenue, no minimum viable product (MVP), low market traction, red flags relating to audit review, data compliance and IP ownership are algorithmically tagged with high to medium risk score due to high perceived business risks. Similarly, startup in the series A-C stage without product validated business models or recurring revenue were flagged under strategic and financial risk. These synthetic rules and domain specific logic were integrated in the dataset to complement historical data and enhance model sensitivity to stage-specific risk factors.

- **Exploratory Data Analysis (EDA) and Unsupervised Profiling:** Statistical and visualisation techniques were deployed for EDA to understand the structure, grouping within the dataset, and validates underlying patterns or profiles in the unlabelled data. The K-Means clustering was applied to segment the startups based on shared characterised across financial, operational, strategic, compliance and technology variables. The Elbow Method was employed to determine the optimal number of clusters (risk levels), evaluating the points at which the marginal gains in the explained variance (inertia) dropped significantly. In addition, the insights from the clustering activities helps to integrate additional features that enhanced the supervised learning model's ability to differentiate business stage and investment readiness.

- **Feature Selection:** To reduce complexity and enhance model efficiency, feature importance techniques such as ***Random Forest classifier, mutual information, correlation filtering using correlation heatmap and the chi-squared test of association*** were applied. This resulting to the final 19 relevant features deployed for the model development.

- **Standardizing Numerical Variables:** Prior to model development, numerical data were normalized using MinMaxScaler to reduce the impact of scale disparities across variables discovered exploratory data analysis (EDA).

### 3.3 Model Building / Predictive Modeling Approach

The study deployed a supervised machine learning strategy in the development of an AI/ML risk intelligence predictive system. This strategy was used to frame startup risk classification (risk score) as the TARGET variable alongside 19 other independent (predictors) variables. To ensure robust model training and unbiased performance evaluation, the dataset was partitioned into three distinct subsets: a training set (70%), a validation set (15%), and a testing set (15%). The training set was used to fit the machine learning algorithms and learn underlying patterns in the data. The validation set was employed for hyperparameter tuning, model selection, and to prevent overfitting. Finally, the test set was reserved exclusively for evaluating the final model's generalisation on unseen data. The stratified splitting approach preserved the distribution of the target variable across subset, ensuring consistency and representativeness of the risk categories. A range of models were benchmarked, including:

- Random Forest Classifier (RF)
- HisGradient Boosting Classifier (HGB)
- K-Nearest Neighbors Classifier (KNN)
- Logistic Regression (LogReg)
- Decision Tree Classifier (DT)

### 3.4 Model Evaluation

The model was evaluated using the confusion matrix drilled down to the standard classification metrics including precision, recall, F1 score, accuracy, and AUC score. In-depth analysis was carried out using visualisations to provide explainable insights into which features most influenced each prediction. Furthermore, to optimize the model performance, the Random Search was employed for hyperparameter tuning during the validation stage. This was selected because it requires less computational resources compared to the Grid Search hyperparameter tuning techniques.

- **Random Search** was deployed due to larger and more complex parameter spaces, allowing efficient exploration by sampling a fixed number of parameter combination at sampling. The above technique was utilized simultaneously with the cross-validation using a validation set to select the best model configuration, thereby minimizing overfitting, and improving generalisation on unseen data.

### 3.5 Model Deployment and User Integration

The most appropriate and best-performing model was deployed as a streamlit web application, allowing real-time interaction for end-user. The application facilitates the ease of consuming and accessing the outcomes of the research as detailed below:

- **Risk Level Classification** (Low, Medium, High).
- **Investment Readiness Assessment** with Explanation.
- **Visualization** comparing user input to industry benchmark for ease of interpretation.
- **Customized targeted recommendations** for improving risk posture and readiness.

These tools aim to support startup founders, investors, accelerators, and policymakers in making informed investment decisions about start funding, support, and scaling strategies in Africa.

| S/N | Factors | Variables Name | Data Type | Note/Explanation |
|---|---|---|---|---|
| 1. | Financials | cash_reserve | continuous/float | Total cash on hand (USD) |
| 2. | Financials | burn_rate | continuous/float | Monthly cash burn (USD) |
| 3. | Financials | ebitda | continuous/float | Earnings before interest, tax, etc |
| 4. | Financials | ebitda_margin | continuous/scale(ratio) | Company's operating income as a percentage of revenue before interest, taxes, depreciation, and amortization |
| 5. | Financials | gross_margin | continuous/scale(ratio) | Remainder of company's revenue after taking out the cost of goods sold (COGS) as a percentage of revenue |
| 6. | Financials | gross_profit | continuous/scale(ratio) | Revenue minus cost of goods sold (COGS) |

| | | | | |
|---|---|---|---|---|
| 7. | Financials | runway_months | continuous/scale(ratio) | Months before cash depletion |
| 8. | Financials | total_asset | continuous/float | Everything the company owns that has values and can generate income (cash, equipment, property, inventory, and receivables) |
| 9. | Financials | total_liabilities | continuous/float | Represents all debts and obligations a company a. company owes to outsiders. |
| 10. | Financials | revenue | continuous/float | Annual revenue (USD) |
| 11. | Financials | funding_stage | ordinal | Bootstrap, Seed, Series A/B/C, or IPO |
| 12. | Financials | fixed_cost | continuous/float | Annual fixed operational costs |
| 13. | Financials | variable_cost | continuous/float | Costs that vary with sales/production |
| 14. | Strategic | customer_concentration | continuous/scale(ratio) | The proportion of a company's revenue that comes from a small number of customers |
| 15. | Strategic | brand_sentiment | continuous/scale(ratio) | Public perception of the business brand - score (0–100) |
| 16. | Strategic | market_traction | continuous/scale(ratio) | Growth and demand for product/service - Score (0–100) |
| 17. | Compliance/Regulatory | IP_ownership | continuous/scale(ratio) | IP ownership strength - Score (0–100) |
| 18. | Compliance/Regulatory | data_compliance | continuous/scale(ratio) | Score (0–100) for data compliance (e.g., GDPR |
| 19. | Compliance/Regulatory | audit_flags | ordinal | Compliance risk score (0 = none, 10 = high risk) |
| 20. | Investment Readiness | risk_score | ordinal | How prepared a business is for investment, based on its risk level. A low-risk score means the business is strong and ready for investment. Whereas a high-risk score means the business is riskier and needs improvement before attracting investors |
| 21. | Operational | cac | continuous/float | Total cash on hand (USD) |
| 22. | Operational | founder_strength_score | continuous/scale(ratio) | Founder experience and track record - Score (0–100) |
| 23. | Operational | team_experience | continuous/scale(ratio) | Skill, knowledge, and track record of a company's leadership and key staff |
| 24. | Operational | customer_churn_rate | continuous/scale(ratio) | Customer loss rate (%) |
| 25. | Operational | employee_count | continuous/scale(ratio) | The total number of people working for a company |
| 26. | Operational | industry | Nominal Categorical Data | Business sector (e.g., Tech, Retail, Health) |
| 27. | Technology | product_stage | ordinal | Current product development stage (e.g., ideation, MVP, Launched) |

**Table 2: Category of Business Risk Factors and Variables**

## 4. Results and Findings

This section presents the results of the research, specifically the model development process, the performance of the different algorithms, and insights derived from the data-driven risk and investment assessment. The model was evaluated based on standard classification success metrics including precision, recall, F1-Score, accuracy, and the interpretability of the model was enhanced using the visualisation and clustering insights.

### 4.1. Model Performance Comparison

| Model | Target | Precision | Recall | F1-Score | Accuracy | AUC Score |
|---|---|---|---|---|---|---|
| Random Forest Classifier (RF) | Low Risk: 0 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| | Medium Risk: 1 | 0.99 | 0.99 | 0.99 | | |
| | High Risk: 2 | 0.99 | 1.00 | 0.99 | | |
| Hisgradient Boosting Classifier (HGB) | Low Risk: 0 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| | Medium Risk: 1 | 0.99 | 0.99 | 0.99 | | |
| | High Risk: 2 | 0.99 | 0.99 | 0.99 | | |
| KNeighbors Classifier (KNN) | Low Risk: 0 | 0.81 | 0.75 | 0.78 | 0.79 | 0.79 |
| | Medium Risk: 1 | 0.73 | 0.76 | 0.75 | | |
| | High Risk: 2 | 0.84 | 0.86 | 0.85 | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Logistic Regression (LogReg) | Low Risk: 0 | 0.75 | 0.71 | 0.73 | 0.73 | 0.73 |
| | Medium Risk: 1 | 0.65 | 0.66 | 0.66 | | |
| | High Risk: 2 | 0.79 | 0.81 | 0.80 | | |
| Decision Tree Classifier (DT) | Low Risk: 0 | 0.98 | 0.99 | 0.99 | 0.98 | 0.98 |
| | Medium Risk: 1 | 0.99 | 0.98 | 0.98 | | |
| | High Risk: 2 | 1.00 | 0.99 | 0.99 | | |

**Table 3: Model Evaluation Metrics on Test Set**

### Key Finding

Random Forest Classifier emerged as the best performing, with an accuracy and AUC score of 99% indicating a strong balance across all metrics, making it ideal for practical deployment in risk classification and readiness prediction tools. Other ensemble method, such as Histogram-based Gradient Boosting (HGB), also demonstrated excellent performance, with a high level of precision, recall, AUC scores, reinforcing the robustness of gradient-boosting techniques in high-stake prediction tasks.

| **Metrics** | **Score** |
|---|---|
| Mean | **0.9877** |
| Std | **0.0008** |
| Best | **0.9886** |
| Worst | **0.9864** |
| **StratifiedKfold (n_split=5) Results** | **[0.98714286, 0.98785714, 0.98642857, 0.98857143, 0.98857143]** |

**Table 4: Cross-Validation Results using StratifiedKfold (n_split=5)**

| **Hyperparameters** | **Declared Dictionary of Hyperparameter Tuning** | **Best Estimator Hyperparameters** |
|---|---|---|
| n_estimators | {50, 100} | 99 |
| max_depth | {3, 20} | 16 |
| min_samples_split | {2, 10} | 5 |
| Min_samples_leaf | {1, 10} | 1 |

**Table 5: Random Search Hyperparameters Tuning**

### 4.1 Feature Importance and Explain Ability

To interpret the predictions and enhance model transparency, Random Forest feature importance, mutual information, correlation matrix and chi-squared statistics for test of association were deployed to assess feature importance across top performing models. The top 20 relevant features are shown in figure 2 below. In addition, Table 6 and 7 below also filtered and expand the relevant features to top 8 and 24 most influential features respectively.

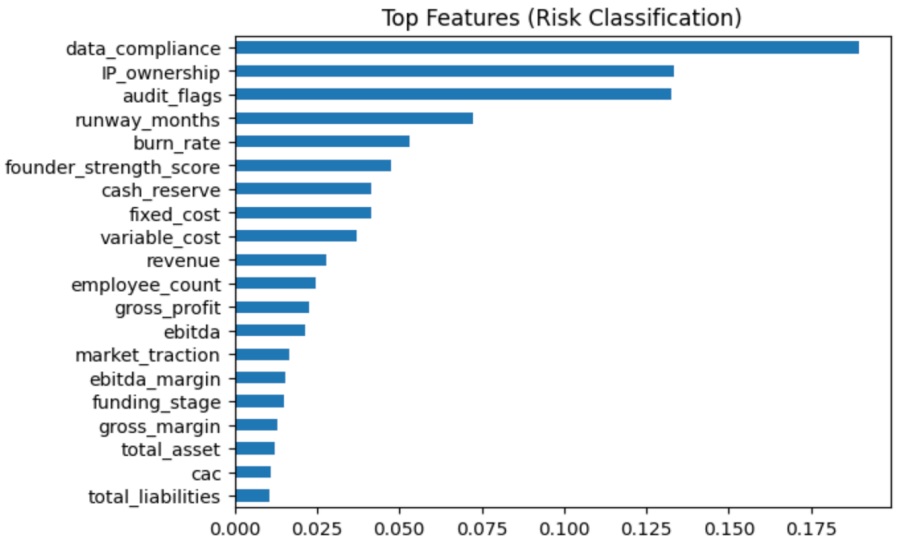


**Figure 2:** Top 20 Relevant features (Risk Classification) using Random Forest Classifier

## Key Finding

Across all the relevant feature selection techniques, features such as data compliance, IP ownership and audit/review flags were strong indicators of compliance and regulatory risks. Whereas, runway (in months), burn rate, cash reserves, fixed cost, variable cost, revenue, gross profit, ebitda, ebitda margin, customer acquisition cost (cac), funding stage, total assets and total liabilities are most influential features for financial risks. Founder strength score and market traction indicators of operational and strategic risks respectively. These most influential features were prioritised in the downstream feature selection and model building and tuning process. The findings align with existing studies, domain expertise, and reinforce the credibility of the model. The overall feature correlation analysis depicted in (Figure 2) show an overall sparsity of strong correlations across features supports the robustness of ensemble and tree-based models, which are less sensitive to multicollinearity, but still benefits from diversified input representations. Furthermore, highly correlated features especially the derived features were carefully evaluated to avoid introducing noise or multicollinearity which can create distortion in interpretability of the model ultimately leading to overfitting.

| Rank | Feature | Chi-Square Score |
|---|---|---|
| 1. | employee_count | 4844.447653 |
| 2. | audit_flags | 3941.235040 |
| 3. | market_traction | 3719.012142 |
| 4. | funding_stage | 3568.197754 |
| 5. | product_stage | 3275.383118 |
| 6. | data_compliance | 1787.011574 |
| 7. | IP_ownership | 1258.664746 |
| 8. | founder_strength_score | 999.672505 |

**Table 6: Top 8 Features for Modeling (Ranked by Chi-Square Statistics): Higher values indicate stronger association with the target variable**

| S/N | Feature | MI (Risk Score) |
|---|---|---|
| 1. | total_asset | 0.000000 |
| 2. | data_compliance | 0.076457 |
| 3. | cac | 0.097892 |
| 4. | audit_flags | 0.205479 |
| 5. | customer_churn_rate | 0.000000 |
| 6. | gross_margin | 0.041974 |
| 7. | variable_cost | 0.219202 |
| 8. | IP_ownership | 0.060776 |
| 9. | brand_sentiment | 0.003862 |
| 10. | total_liabilities | 0.005538 |
| 11. | team_experience | 0.000000 |
| 12. | runway_months | 0.043330 |
| 13. | customer_concentration | 0.000000 |
| 14. | funding_stage | 0.211405 |
| 15. | industry | 0.000000 |
| 16. | gross_profit | 0.203514 |
| 17. | revenue | 0.214520 |
| 18. | ebitda | 0.207378 |
| 19. | burn_rate | 0.224555 |
| 20. | founder_strength_score | 0.054249 |
| 21. | cash_reserve | 0.235765 |
| 22. | market_traction | 0.213163 |

| | | |
|---|---|---|
| 23. | product_stage | 0.202308 |
| 24. | fixed_cost | 0.223586 |

**Table 7: Top 24 features for modeling (Ranked by Chi-Square Statistics): Higher values indicate stronger association with the target variable**

Figure 3 below presents the feature correlation matrix in heatmap form to identify multicollinearity and interdependency amongst predictor variables. The analysis reveals that while several features- especially the most influential features show moderate to strong correlation (r2>70%), the remainder of the features are either weakly correlated or independent. Table 8 evaluates the strength of the association between input features and the target variable. A chi-squared test of association (Table 8) was implemented and revealed that subset of the features demonstrated statistically significant relationships with the target variable ($p<0.05$) suggesting these features are carry significant predictive capability. Notably, features such as data compliance, audit flags, funding stage, intellectual property (IP) ownership, variable costs, founders' strength score in terms of leadership experience, skills and track record exhibit the strongest associations, including the probability that their inclusion in the model will contribute positively to the model performance.

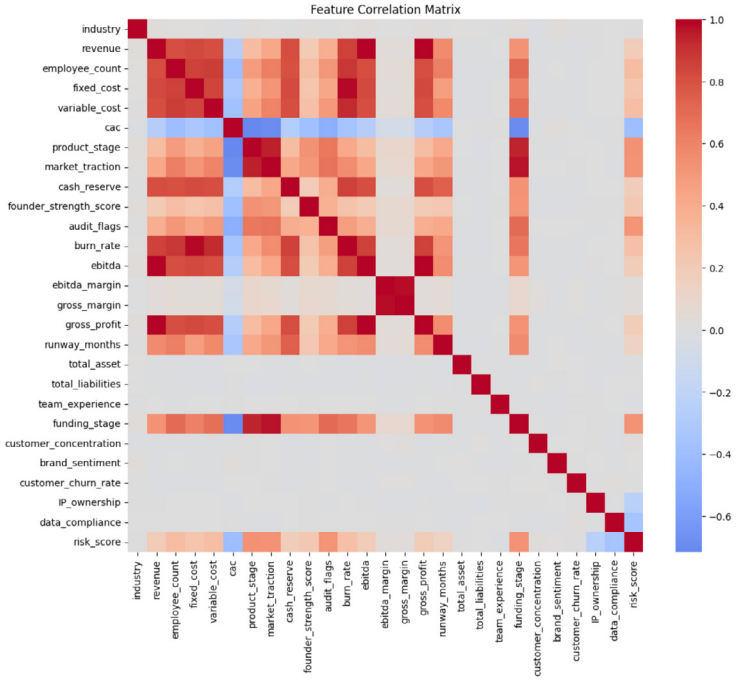


**Figure 3:** Feature Correlation Matrix

| S/N | Feature | Chi2 | p-value | Decision Rule: (Significant Associated with the Target Variable if p-value < 0.05) |
|---|---|---|---|---|
| 1. | total_asset | 19898.146584 | 5.063725e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 2. | *data_compliance | 1787.011574 | 1.733516e-253 | Significant Feature (**p-value < 0.05)** |
| 3. | cac | 20000.000000 | 4.946808e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 4. | *audit_flags | 3941.235040 | 0.000000e+00 | Significant Feature (**p-value < 0.05)** |
| 5. | customer_churn_rate | 171.977247 | 9.092136e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 6. | gross_margin | 20000.000000 | 4.946808e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 7. | *variable_cost | 18734.667437 | 9.306802e-02 | Significant Feature (**p-value < 0.05)** |
| 8. | *IP_ownership | 1258.664746 | 1.184856e-153 | Significant Feature (**p-value < 0.05)** |
| 9. | brand_sentiment | 216.639038 | 1.730898e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 10. | total_liabilities | 19887.903520 | 4.868591e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |

| | | | | |
|---|---|---|---|---|
| 11. | team_experience | 200.787540 | 4.313660e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 12. | runway_months | 20000.000000 | 4.946808e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 13. | customer_concentration | 190.690926 | 6.324317e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 14. | *funding_stage | 3568.197754 | 0.000000e+00 | Significant Feature (**p-value < 0.05)** |
| 15. | industry | 18.804427 | 4.039586e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 16. | gross_profit | 20000.000000 | 4.946808e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 17. | revenue | 20000.000000 | 4.946808e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 18. | ebitda | 19996.732760 | 4.972083e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 19. | burn_rate | 19478.256322 | 2.598655e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 20. | *founder_strength_score | 999.672505 | 2.203883e-206 | Significant Feature (**p-value < 0.05)** |
| 21. | cash_reserve | 20000.000000 | 4.946808e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 22. | *market_traction | 3719.012142 | 0.000000e+00 | Significant Feature (**p-value < 0.05)** |
| 23. | *product_stage | 3275.383118 | 0.000000e+00 | Significant Feature (**p-value < 0.05)** |
| 24. | fixed_cost | 19382.154533 | 2.900529e-01 | Non-Significant Feature (**p-value ≥ 0.05)** |
| 25. | *employee_count | 4844.447653 | 4.006718e-270 | Significant Feature (**p-value < 0.05)** |

**Table 8: Test of Significant Association with Target Variable using Chi-Square**

*** Means Statistically Significant Association With the Target Variable**

### 4.2 Risk Profile Clustering

As depicted in (Table 9) and (Figure 4) below, K-means clustering supported by the elbow method was utilized to segment each instance of the data (representing) startups into their natural grouping based on risk profiles and select the optimal number of clusters as shown in the below Figure 4. These clustering techniques offer nuanced insights beyond the binary risk classifications, supporting investors, technology hub/accelerators in making informed investment decisions.

| Cluster Summary | Risk Score | Startup Risk Profile and Investment Readiness | Features | Funding Stage/Business Development Lifecycle |
|---|---|---|---|---|
| Cluster 1 | Risk Score: 0 | Low Risk, High Readiness | Mature product, decent founders experience and track record, strong financial governance, stable cashflow, strong data compliance, intellectual property (IP), and market traction | Series C and beyond, IPO |
| Cluster 2 | Risk Score: 1 | Medium Risk, Emerging Readiness | Early Market/proven business model but limited financial buffers, moderate founders experience and track record, relatively good data compliance, intellectual property (IP) and market traction, reasonable/early users. Investors are usually VCs. | Series A, Series B |
| Cluster 3 | Risk Score: 2 | High Risk, Low Readiness | Little or no revenue (pre-revenue), high burn rate, governance gaps, early-stage product lifecycle, limited founders experience and market track record, high failure rates. Funding sources from founders, friends/family, or angel investors. | Bootstrapping, Idea/Seed Stage |

**Table 9: Risk Profile Clustering using Unsupervised ML method: K-Mean and Elbow Method**

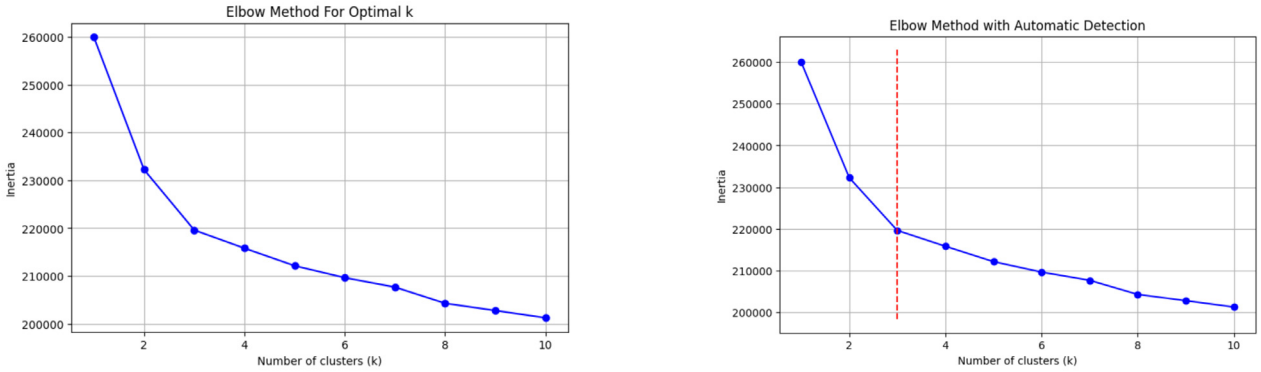


**Figure 4:** K-Mean and Elbow Method for Optimal Cluster (k=3)

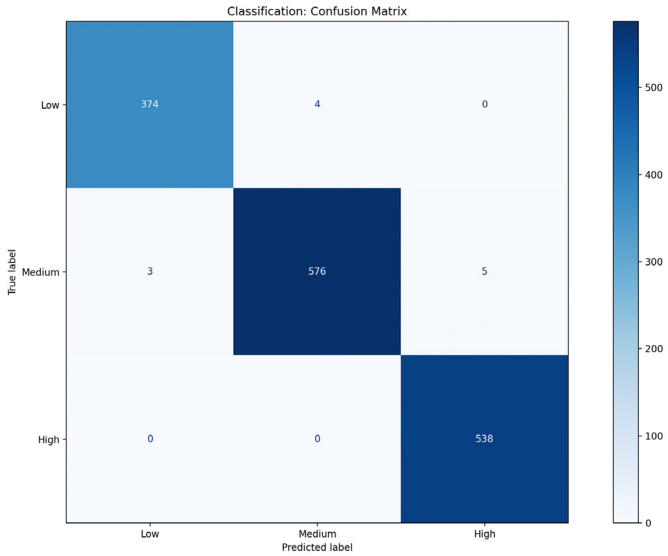


**Figure 5:** Classification: Confusion Matrix Report

The confusion matrix for the Random Forest Classifier (Figure 5) displayed a high predictive accuracy with 374/377 for risk class: 0, 576/580 risk class: 1, and 538/543 for risk class: 2 test set correctly classified. The model completely avoids flagging a high promising, investable startup as high risk when it is not (Type 1 Error - False Positive), it also avoids flagging a high-risk startup or approving it as investment-ready (low/medium risk) – Type 2 Error – False Negative.

### 4.4. Model Deployment

To facilitate open access and availability of the research for the targeted audiences, the trained model was deployed using a Python-based Streamlit interface, enabling a real-time risk intelligence and investment readiness assessment tool. A QR code was embedded to allow mobile access and customised downloadable PDF report of model explanations and risk diagnostics.

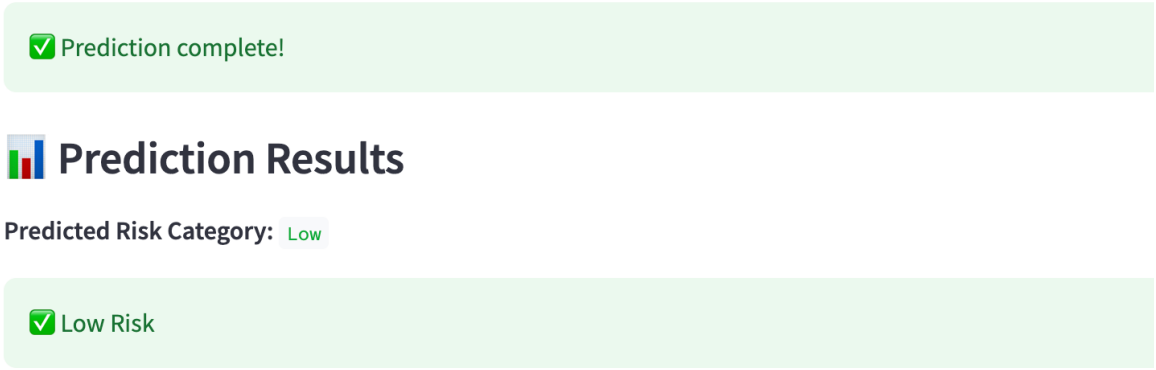


**Figure 5.1 :** Scan QR code and Prediction Output of AI-Powered Risk Intelligence for Investment Readiness

## 5. Discussions and Interpretation of Findings

This study explored the application of artificial intelligence and machine learning to assess startup risks and investment readiness within the unique context of African Innovation ecosystems. The findings provide critical insights not only into the technical performance of the predictive models but also into the broader implications for startups development, investor decision-making, and policy supports in emerging markets such as Africa.

### 5.1 Bridging the Information Gaps in Investment Decisions

As detailed in Table 3, 4 and 5 above, the predictive models demonstrated strong performance, particularly the Random Forest (RF), Hisgradient Boosting Classifier (HGB) and Decision Tree (DT) Classifiers which achieved accuracy score of 99%, 99% and 98% respectively. The strong performance was based on a balanced dataset – classes evenly distributed, supplemented with a robust validation resulting from a cross-validation (CV) on the validation set (to avoid any data leakage). The CV produced an average, best and worst CV of 98%. By adopting a multi-dimensional and data-driven assessment methodology of risk across different aspects of the business such as financial, operational, compliance/regulatory, technology and strategic dimension, the model helps bridge the information gaps which is one of the biggest challenges for startup funding in Africa. Furthermore, the inclusion of real world startup dynamics such as funding stage, product development lifecycle and governance readiness ensure that the model reflects the nuanced challenges that founders and investors navigate across the African innovation ecosystem.

### 5.2 Consolidation of the Power of Feature Engineering and Domain Expertise in Finance

The study strength and its comparative advantage lies in the integration of domain expertise into the data simulation and feature engineering process. At the individual feature level, features such as burn rate, revenue, product development stage, runway, gross margin, ebitda might not have been statistically significant at the individual, but their interactions with other features such as market traction, funding stage, data compliance, IP ownership, founder strength score make them very influential and highly intuitive features. Hence, this alignment enhances the trust in the model and demonstrates how domain informed AI design can yield more meaningful insights for diverse stakeholders. More importantly, the application of rule-based logic to stimulate risk profile at various funding and product development lifecycle ensured that the model retained contextual relevance, despite the synthetic nature parts of the dataset. This methodology rigor strengthens the practicality and real-world application of the study's contribution to applied AI in underserved data environments.

### 5.3 Enhancing Explainability and Decision Confidence

The use of random forest classifier feature importance, mutual information, Chi-squared test of association and correlation heatmap as depicted in Figure 2, Figure 3, Table 6 and 7 provided transparency into the features selection process, decision making and highlight the most influential features - how the different features interact to influence the predictive capability of the model. In addition, the transparency in the feature selection process enhance trust needed for adoption of the model. As depicted in Table 9, rather than presenting a 'black box', the model reveal why startups are clustered in different group (high risk, low readiness; medium risk, emerging readiness; or low, high readiness) based on common profiles and unique business context. This interpretability enhances the confidence of investors, founders, and support organisations who can act on specific insights such as improving financial runways and resolving compliance issues. The confusion matrix report in (Figure 5) demonstrated very small error across risk classes which confirms outstanding predictive performance of the Random Forest Classifier Furthermore, the use of the K-means and elbow method clustering techniques allows the identification of distinct startups grouping beyond binary classification, the cluster reflect real-world diversity in startup readiness, enabling stakeholders to tailor their interventions based on specific startups profiles.

### 5.4 Limitation and Considerations for Practical Application

Due to the absence of publicly available, multi-dimensional datasets for technology startups, synthetic data was generated and combined with small-scale real-world cases based on domain knowledge and industry benchmarks. Although the model demonstrated strong performance, its reliance on simulated dataset with domain knowledge introduces some challenges. While data is synthetic, the model predictions were validated against known

behaviours patterns of technology businesses, corroborated by expert inputs. In addition, features utilized were synthesized using statistical techniques derived from prior market research, insights from several incubation and acceleration programmes over the last 5 years and validated with domain experts to stimulate realistic startups users behaviours. Furthermore, the qualitative features such as founders' strength score, IP ownership, data compliance, – although encoded, but remains challenging to model purely through structured data. Hence, a mixed model that integrates NLP from pitch decks, financial, legal, and commercial documents, face-to-face interviews with founders may improve holistic assessments. This research lays the groundwork for future practical applications and continual learning framework and validations as more live startups data is acquired and anonymised from accelerators, VCs, enterprise support organisations, and government entrepreneurship programmes to enhance practical generalization.

### 5.5 Contribution to Knowledge and Research Gaps

This study contributes to existing literature by offering a practical framework for AI/machine learning powered risk intelligence and investment readiness scoring in a region often overlooked in global datasets. In addition, it addresses gaps in previous studies models that were limited to financial metrics, non AI/ML driven model, or lacked explainability. By integrating domain knowledge, simulation, rule- based logic and explainability, this research offers a methodological innovation backed with actionable tools for a diverse stakeholder in the African innovation economy.

## 6. Conclusion and Recommendations

The overarching aim of this research was to develop and validate a machine learning framework for assessing startup risk and investment readiness, tailored specifically to the African innovation ecosystems. Consistent with this aim, the study utilized a hybrid approach integrating simulated and expert informed datasets to develop a robust AI/ML model capable of categorizing startups across critical risk dimensions including financial, operational, compliance, technology and strategic. The findings from the research highlights the importance of methodological innovation and the transformative potential of artificial intelligence in uncovering the business risks associated with high growth, technology ventures and driving informed investment decisions for productive allocation and access to capital. Furthermore, the research leveraged explainable models and contextualised features - making the research to be practically relevance and provides a framework that outperformed traditional generic risk scoring. It also integrated the unique and nuanced realities of early-stage startups in the continent. The methodological innovation of the research birthed a framework that bridge the information gap between startups and investors.

### 6.1 Practical Implications for African Startup Ecosystem

The outcomes of this research have several immediate relevance to diverse stakeholders in the startups and investment ecosystem – including but not limited to startup accelerators, angel investors, venture capital (VCs), and government programmes and interventions aiming to support entrepreneurship. The research offers a standardized, scalable, and explainable risk intelligence and assessment tools as explained below:

- **Startups and Founders:** The model provides investment readiness diagnostics platform for early detection of business risk. It also provides actionable insights and mitigants to close the identified risks. Startups can have full visibility into the factors influencing their investment readiness resulting from the different dimensions of the business.
- **Accelerators, Incubators and Funded Programs:** The AI/ML predictive tool can support the design and implementation of an inclusive and tailored capacity building initiatives in different acceleration and incubation programs. Furthermore, it can drive cohort selection, progress tracking and impact reporting, ultimately serving as a pre- and post-diagnostic tool for investment supports.
- **Investors and Venture Capital:** The model enables due diligence and pipeline qualification for startup accelerators. In addition, the model offers a data-driven and consistent screening process for startup applications which can ultimately reduce due diligence time while enhancing decision confidence.
- **Policy and Development Agencies:** The model provides evidence-based inputs for design and implementation of policy in enterprise development programs, access the health of innovation ecosystems and funding allocations.
- **Academic and Higher Education Institutions:** The outcomes and dataset from this research provides open access for the future research work in AI/ML, the development of robust and more generalisable predictive models for the African innovation ecosystem. In addition to the above immediate relevance of the research, the deployment of the model through a web-based application enhances accessibility, allowing ecosystem stakeholders to leverage the insights without technical expertise.
- **Policy Recommendations**

  The below policy actions are recommended to harness the full potential of artificial intelligence and machine learning in startup development and data-driven investment decisions:
- **AI-Driven Investment Readiness and Governance Infrastructure:** There is an urgent need for public and private partnerships to co-invest in platforms that supports standardized risk diagnostics, embedded in accelerators and funders ecosystems.
- **Development of Local AI Talent with Sectoral Focus:** Given that AI/ML are emerging technologies that requires heavy domain expertise to build a contextual-aware and practical predictive models. Hence, investment in AI education should align with enterprise development needs, specifically in financial services, legal tech, agriculture, education, SME support systems amongst others.
- **Open Access and Democratisation of Startup Data:** Driven by appropriate privacy controls, government, and enterprise support organisations (ESO) should encourage data sharing policies that enables AI the development of better generalisable AI modeling at scale.
- **Ethical and Inclusive AI:** Despite the myriad benefits and

the capability of AI, there is still a big need for AI governance framework that ensure transparency, equity, and explainability in decision-making.

The confusion matrix (Figure 5) confirmed the model reliability, with minimal misclassification across all categorises. The Random Forest (RF) model not only matched but exceeded expectations in classifying multi-class risk categories with an exceptional recall and precision score. With only minor errors out of 1,500 predictions, and an AUC score of 99%. The consistency of the outputs makes its particularly well suited for domains where decision accuracy is critical. Based on the above findings, practical implications and policy recommendations, the research not only provide a technical solution but also offers a practical framework for scaling smarter funding decisions, de-risking innovative ventures, and enhancing enterprise development in Africa.

## Summary Table

The section summarizes what currently exists and how this AI/ML research contributes to the body of knowledge to startup risks and investment readiness.

| Aspects | What Already Exists | How this research contributes to the body of knowledge |
|---|---|---|
| **Problem Area** | Startup risk evaluation often rely heavily on professional judgment, expert scoring, or historical financial information. This is prone to human error and very time consulting – can take between 10-24 weeks. | This research offers a scalable, dynamic, and repeatable machine learning-decision support system for objective and timely risk and readiness prediction. |
| **Utilization of ML techniques** | Existing studies utilized theoretical framework and traditional methods such as linear regression which may not fully captured the nuanced/complex drivers of startup risk and business contexts. | This research applied and benchmarked five different advanced ML algorithms including two ensemble methods for multi-class risk classification including: Random Forest, HistGradient Boosting, Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN). |
| **Conservativeness and risk-averseness nature of the model** | Many traditional models treat false positives (false alarm) and false negatives (miss) equally | Prioritizes recall as the most influential evaluation metric especially in the high- |
| | – resulting from over reliance of 'accuracy' as the sole evaluation metric. | risk class - to minimize Type II errors (false negatives: missing a risky startup). This aligns with the risk-averse investors behaviour. |
| **Sector-specific Relevance** | Most research is generic, focused on developed market with limited considerations for the African innovation ecosystems. | This research tailored its methodology and assumptions for African innovation ecosystems, contributing context-aware AI to under researched region. |
| **Evaluation/Success Metrics** | Previous studies focused on accuracy or overall performance only. | Utilized a more robust and context-aware evaluation metrics such as confusion matric analysis: class-level recall, precision, and AUC, specifically optimising for low Type II error. |
| **Interpretability and Trust** | The model utilized in prior research are often treated as 'black boxes' with limited explanation. | This research used several explainable techniques such as the correlation feature heatmap, Random Forest features selection, mutual information for the selection of relevant features. It also show how each feature increase or decrease the predictive power of the model. Furthermore, it incorporates class-wise performance interpretation, enabling actionable insights for investment decision and regulatory alignments. |
| **Data Use** | Relied heavily on qualitative analysis or post funding performance and limited use of structured data. | Utilized structured, labelled startup data to train classification models that predicts risk category throughout the business life cycle. |
| **Usability and Development** | Research is focused on mostly academic, not user-focused or field deployable. | Design for real-world use by integrating streamlit app and downloadable PDFs, QR code access, visual dashboards for startup founders and fund managers. |

| | | |
|---|---|---|
| **Practical Applicability and Impacts** | Limited applicability and implementation of ML in venture screening workflows. | This research demonstrated feasibility of deployable predictive systems that can flag high-risk startups before capital is committed. |


## Acknowledgements

I would like to express my sincere gratitude to Co-Creation Hub Limited (CcHUB) for providing the platform to experience real-world startup supports and data that have made this work successful. I am also grateful to the team at Gauge Innovation Limited for providing the collaborative environment for constructive discussions, reviews by domain experts, support during the development, and testing of the ML predictive models.

I also acknowledge the open-source community and the developers of the tools and libraries that were instrumental in the implementation of the models.

Finally, I thank my family and friends for their unwavering support and patience during this research journey.


## Data Availability Statement

The project dataset analysed during the investigation phase of this study was simulated from a small- scale real-world scenario. Due to the proprietary nature of the real-world data and the confidentiality agreements with participating organizations, the dataset is not publicly available. However, simulation code and rule-based logic files used in generating the 10,000 synthetic datasets utilized for training and developing the ML models are available upon reasonable request to the corresponding author.

## Ethical Statement

Domain experts who contributed to this study were fully informed of the research goals and provided explicit consent for their knowledge to be used in this work and were given the opportunity to withdraw from the research at any time, without reason. In addition, the study was conducted in accordance with the ethical standards outlines by the participating organisations. Efforts were made to minimize bias in both real-world data collection and simulation processes to ensure fair and equitable outcomes. The study did not engage any less privileged, physically challenged individuals nor animal subjects.

## Ethics Approval

The actual dataset utilized in this research to train, validate, test, and develop the machine learning models were generated using rule-based and simulation techniques to model real-world scenarios. This contains no personally identifiable information and completely anonymized. However, Ethical approval was applied and granted for this research by Gauge Innovation Limited.

## Conflict of Interest Statement

The author declares no conflicts of interest related to this research. This work was conducted independently and was not influenced by any commercial, financial, or organizational relationships that could be construed as a potential conflict of interest.

## References


1. Nobel, C. (2011). *Why companies fail--and how their founders can bounce back*. Boston, MA: Harvard Business School.
2. Church, J. (2020). Investable entrepreneur: How to Convince Investors Your Business is the Best One to Back. *1st ed., UK: Rethink Press. pp.49*.
3. The Institute of Chartered Accountants of Nigeria (ICAN). (2021). *Performance Management*, 2021. Nigeria: *The Institute of Chartered Accountants of Nigeria*, pp.627-660.
4. Barber, J. (2020). Investor Ready: The guide for start-ups on getting investors to say Yes. 1st ed., *UK: Rethink Press*. pp.263-276.
5. Russell, S. and Norvig, P. (2021). Artificial intelligence: A Modern Approach. *4th ed., Harlow, UK: Pearson Education Limited*.
6. Hopper, G. (2021). *Deep Finance: Corporate Finance in the Information Age*. 1st ed. USA: Leaders Press,. pp.25-51.
7. Mitchell, T. M. (1997). *Machine Learning*, 1st ed., New York, NY: McGraw Hill.
8. Tech Cruch (2024). *How African Startups Raised Funding in 2023*.
9. Aziz, S., Dowling, M. (2024). "AI and machine learning for risk management", *Social Sciences Research Network (SSRN)*.
10. Laseter, T., Frazer, A. Boatright, B. (2022). "Artificial intelligence in business: machine and management", S*ocial Sciences Research Network (SSRN)*.
11. Petersons, E. (2024). "Artificial intelligence (AI) and machine learning in digitalization", *Social Sciences Research Network (SSRN)*.
12. Abikoye, B. (2021). "Machine learning models and AI for predicting financial crises: applications and accuracy", *Social Sciences Research Network (SSRN)*.
13. Krause, D. (2023). "ChatGPT and other AI models as a due diligence tool: benefits and limitations for private firm investment analysis", *Social Sciences Research Network (SSRN)*.
14. Sanz-Prieto, I., De-la-fuente-Valentín, L., Ríos-Aguilar, R. (2021). "Technical due diligence as a methodology for assessing risks in start-up ecosystems: an advanced approach", *Information Processing and Management, 58*(5).
15. Malhotra, Y. (201.8). "AI, Machine learning & deep learning risk management & controls: Beyond deep learning and generative adversarial networks: Model risk management in AI, machine learning & deep Learning: Princeton presentations in AI-ML Risk Management & Control Systems (Presentation Slides)", Princeton presentations in AI & machine learning risk management & control systems, Princeton Fintech & Quant Conference, *Princeton University*.

16. Zhao, J., Li, B. (2022). “Credit risk assessment of small and medium-sized enterprises in supply chain finance based on SVM and BP neural network”, *Neural Computing & Applications, 34*, 12467–12478.

17. Agbo, C. G. (2022). “Predicting startup status using funding and sentimental data”, *Research Master thesis, Solent University, Faculty of Business, Law, and Digital Technologies, Southampton, UK*.