Random forest model for predicting business success for micro, small and medium enterprises in Kakamega County, Kenya

Sandra Khagayi[1*]
Collins Odoyo[2]
Dorothy Rambim[3]

[1*]sandrakhagayi@gmail.com
[2]codoyo@mmust.ac.ke
[3]drambim@mmusta.c.ke

[1,2,3]Masinde Muliro University of Science and Technology, [1,2,3]Kenya

https://doi.org/10.51867/ajernet.6.3.24

ABSTRACT

The Micro, Small, and Medium Enterprises (MSMEs) are an important part of the economy of Kenya, yet they have a high rate of uncertainty and failure because of complicated, poorly comprehended reasons. This paper has come up with a machine learning model based on the Random Forest algorithm to forecast the success of MSMEs in Kakamega County, Kenya, based on historical data provided by the county government. The study was guided by resource-based view theory which posits that a business's success and sustainability hinge on its ability to acquire, control and utilize valuable internal resources. The traditional approaches that limit themselves to linear financial indicators have been inadequate in describing the multidimensional risks experienced by MSMEs. The study adopted simulation research design to develop a predictive model based on the Random Forest algorithm to predict the success of MSMEs. Random Forest model has shown outstanding predictive accuracy with a precision of 99.72 percent in predicting the success of businesses. The major predictors were found to be the availability of financial access, business characteristics and government support factors. A binary logistic regression model was also used to confirm the results and explained 99.64 percent of the variance in business outcomes. The findings provide a solid basis of evidence-based policy-making and interventions in support. The research is an addition to the existing evidence on the applicability of machine learning in enterprise sustainability and offers a scalable solution to enhancing the resilience of MSMEs in comparable settings. The study successfully developed highly accurate machine learning model for predicting MSME success. It was able to identify critical factors for MSME success, financial access and government support. The study recommends that policy makers and stake-holders should utilize data-driven insights for targeted interventions to enhance MSME resilience and growth. To foster growth and development, MSMEs are advised to focus on improving financial management and leveraging government support programs.

Key words: Business Success, Machine Learning, MSMEs, Predictive Modeling, Random Forest Algorithm

I. INTRODUCTION

Micro, Small, and Medium Enterprises (MSMEs) are a group of businesses categorized in size based on their nature and number of employees. These enterprises play an important role in the economy of Kenya, as they make up 33.8 percent of the national gross domestic product (GDP) and employ more than 80 percent of the population, in accordance to reports from Kenya National Bureau of Statistics (Kenya National Bureau of Statistics [KNBS], 2022). Although MSMEs are very crucial, they experience high failure rates, which are mainly attributed to unpredictable operating environments, absence of formal structure, and inaccessibility to critical resources. In counties such as Kakamega where most livelihoods rely on MSMEs, high death rate of these businesses has been a concern to policymakers, financiers and development stakeholders.

Conventional business evaluation techniques like financial statement analysis, credit scoring and business plan reviews are ineffective in forecasting MSME results because they are based on structured information and linear assumptions (Gichuki et al., 2014; Beck and Demirguc-Kunt, 2006). These models do not pay attention to non-financial indicators and are not able to reflect the complex, non-linear relations that determine business success especially in informal and data-poor environments. This means that the stakeholders cannot act proactively and many MSMEs are exposed to market shocks, inefficient resource allocation and uneven growth.

Random Forests Machine learning models have been proven effective in developed countries, to assess the performance of SMEs and predict their business outcomes. In Germany, a study examined a sample of approximately three million SMEs, both financially and non-financially, through Random Forest. The analysis demonstrated that Random Forests performed better than the conventional approaches in defaulting of SMEs, indicating that they are

useful in the credit risk analysis of such data (Siggelkow & Fernandez, 2024). This success demonstrates that machine learning can be used to play a central role in evaluating and supporting SMEs in regions where more data are available.

This paper addresses these issues by employing the Random Forest machine learning algorithm to create a predictive model that will accurately classify MSMEs based on their chances of success. The model combines important predictors of financial access, business characteristics, infrastructure, market conditions, and government support factors using secondary data of the Kakamega County Government. Random Forest was chosen due to its effectiveness in handling noisy, incomplete and unstructured data and the capacity to identify complicated variable interactions without relying on linearity (Breiman, 2001). Besides increasing the accuracy of prediction, the algorithm will also determine the most factors that influence the success of MSMEs. The model offers a decision support tool to policymakers, financiers, and entrepreneurs, which is data-driven and allows targeted and proactive interventions to enhance the viability of businesses at the county level and beyond.

1.1 Statement of the Problem

The unstable conditions that MSMEs in Kenya experience keep them from getting helpful support and from growing sustainably. Because MSMEs are so informal, flexible and complicated, traditional models of performance measurement fail to capture the real picture and result in poor decisions. For example, Kinyanjui (2014) found that traditional assessment tools overlook the informal practices and irregular income flows common among Kenyan MSMEs, thereby misclassifying many viable businesses at high risk. Similarly, Mwania and Muganda (2019) highlighted that traditional assessment models often assume linear relationships and complete data which rarely exists in the context of MSMEs resulting in poor predictive power. These challenge underscores the need for an advanced, flexible and data-driven methods. Machine learning algorithms like Random Forests have demonstrated ability to handle noisy, incomplete and non-linear data hence are better suited for capturing the complex determinants of MSMEs success. Although such tools are promising, their use in MSME research has not been widespread which creates a big gap in developing accurate, data-based models for policymaking.

1.2 Research Objective

To develop a predictive model for accurately assessing business success for MSMEs using Random Forest Algorithm

II. LITERATURE REVIEW

2.1 Theoretical Review

The study is based on the Resource-Based View (RBV) theory which posits that success and sustainability of a business largely depends on its ability to acquire, control and utilize valuable internal resources (Barney, 1991). RBV helps in identifying which factors of MSMEs should be included as features (predictors) in the Random Forest Model. These features represent the tangible and intangible resources and capabilities that according to RBV, contribute to a competitive advantage and success. The study highlights resources such as financial access, government support, and business characteristics as important predictors of MSME success. RBV provides the theoretical justification for why these specific elements are chosen as inputs for the predictive model.

The RBV theory guides the model in interpreting results (feature importance). After the random forest model is developed and trained, it can determine the relative importance of the input factors in predicting MSME success. RBV provides the theoretical lens through which these feature importance scores can be interpreted. Resources that the model identifies as the highly influential on success can be understood as valuable, rare and difficult to imitate assets that can confer a competitive advantage to MSMEs. For instance, if the model identified financial access as a strong predictor, RBV would explain this as a critical, valuable resource for an MSMEs ability to expand and withstand economic shocks.

2.2 Empirical Review

The unstable conditions that MSMEs in Kenya experience keep them from getting helpful support and from growing sustainably. Because MSMEs are so informal, flexible and complicated, traditional models of performance measurement fail to capture the real picture and result in poor decisions. Most of the traditional methods for assessing MSME performance in Kenya use descriptive statistics, analysis of financial statements, credit scoring and reviews of business plans (Gichuki et al., 2017). These approaches overlook the informal practices and irregular income flows common among Kenyan MSMEs, thereby misclassifying many viable businesses at high risk. Traditional approaches expect businesses to have reliable financial records, but many small and informal businesses do not keep such records (De Mel et al., 2009). Second, these models are linear and cannot represent the complex and nonlinear connections

between business performance variables (Berger & Udell, 2006). As most MSMEs lack have a long credit history, they are often not considered by financial institutions, even if they are well-managed (Beck & Demirgüç-Kunt, 2006). Assessments from business plan reviews are likely to prefer entrepreneurs with formal training and may not see the value in those who are resilient and innovative in practical ways.

Due to these shortcomings, it is difficult to determine whether an MSME will survive. Studies have shown that using predictive models based on artificial intelligence and machine learning can improve the accuracy of insights (Hastie et al., 2009). Random Forest Algorithm is a suitable choice for this study as they have been proven effective in assessing performance of MSMEs in developed countries. The algorithm does not depend on strict ideas about data structure or relationships and can find complex patterns that other methods often fail to notice (Cutler et al., 2007).Random Forest is known to handle noisy and unorganized data which is a common characteristic of MSME data (Breiman, 2001; Cutler et al., 2007).

III. METHODOLOGY

The research design adopted in this study was simulation-based research design to come up with a predictive model based on the Random Forest algorithm to predict the success of MSMEs in Kakamega County. A secondary data of 56,623 records of MSMEs, obtained by Kakamega county registries, revenue offices and microfinance institutions. The data recorded important indicators like the size of business, access to finances, tax compliance, access to infrastructure, government support and other business characteristics. The businesses were ranked as micro, small or medium and were ranked as successful or unsuccessful according to the tax compliance and the activity status.

The data was preprocessed by cleaning, encoding, and feature engineering and the businesses were classified as success or failure depending on tax compliance. The model was implemented in Python programming language with the help of libraries like Pandas to manipulate data, NumPy to perform calculations, Scikit-learn to implement RF algorithm, Matplotlib and Seaborn to visualize data. The data was divided into training (80%) and test (20%) set. It used the RandomForestClassifier of Scikit-learn to train the model. The measures of model evaluation were classification accuracy, confusion matrix, ROC-AUC core, and feature importance analysis.

Binary logistic regression was used as the comparative model to determine the statistical significance and confirm Random Forest results. The logistic regression allowed estimating the effects of each predictor by coefficients, p-values, and odds ratios. The concomitant application of the two models provided a strong assessment and valid findings. Relevant authorities were consulted to provide ethical approvals and all data were anonymized to guarantee confidentiality.

IV. FINDINGS & DISCUSSION

4.1 Random Forest Results

Random Forest algorithm has shown an outstanding performance in predicting the success of MSME based on the past data. It had a predictive accuracy of 99.72%. A heatmap visualization of the confusion matrix showed a strong alignment along the diagonal, further confirming the model’s precision in classifying MSMEs as either successful or unsuccessful. The Receiver Operating Characteristic (ROC) curve and its corresponding Area Under the Curve (AUC) metric yielded a value close to 1.0. This near-perfect AUC score strongly confirms the model's excellent discriminative power in distinguishing between successful and failing businesses.

The feature importance output from the Random Forest model in the study shows which factors had the most influence in predicting the success or failure of MSMEs. According to the model, features such as "Financial Access," "Business Characteristics" and Government Support factors emerged as the top contributors to determining whether a business would succeed. This means that MSMEs with better access to finance, strong internal business practices and compliance with government regulations have a higher chance of being successful. These findings emphasize the critical role of financial support, government support and strategic business characteristics in determining success of small businesses. On the other hand, features like "Infrastructural access" and "Market factors" were relatively less influential in the model. This does not mean they are unimportant, but in comparison to the other factors, they played a smaller role in distinguishing between successful and unsuccessful businesses. This suggests that either government support is limited or not well-targeted, or that businesses have found ways to succeed despite infrastructural challenges. Therefore, while all factors matter to some extent, policy efforts aiming to improve MSME performance should prioritize enhancing financial access and government support programs in order to foster SME success. These findings indicate the ability of the algorithm in capturing nonlinear, complex relationships among variables, which makes it appropriate in unstructured or semi-structured MSME data environments.

```
#Random Forest
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy * 100:.2f}%')
```

```
Accuracy: 99.72%
```

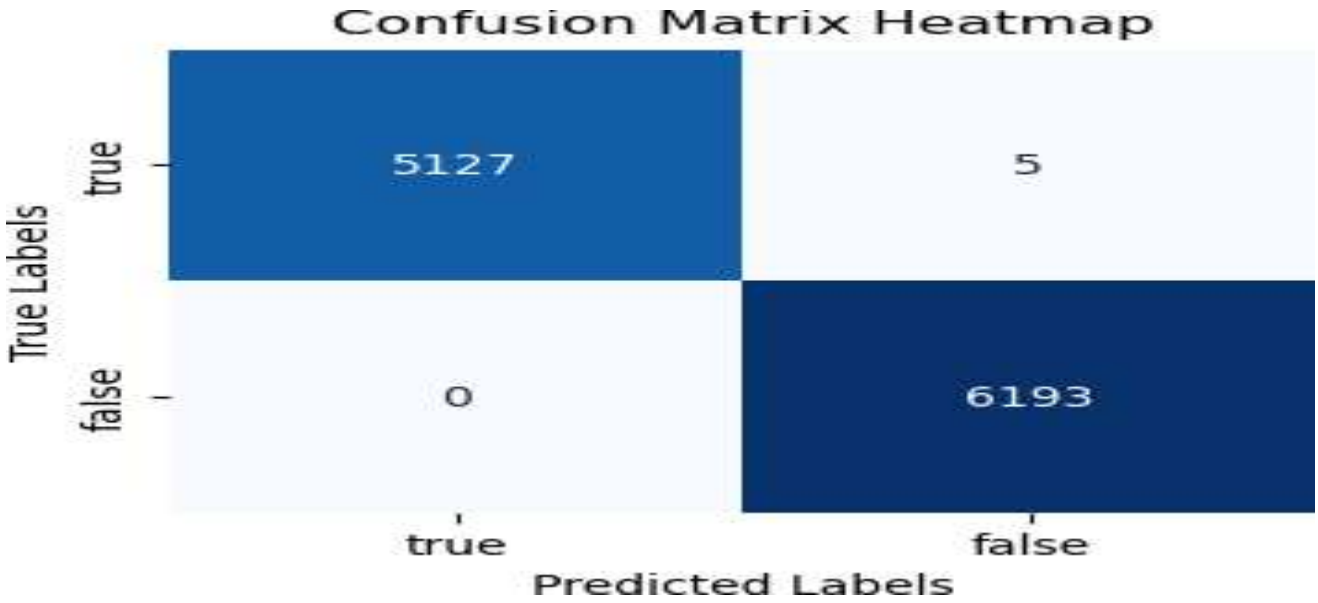

Figure 1
Random Forest Model Confusion Matrix of MSME Success

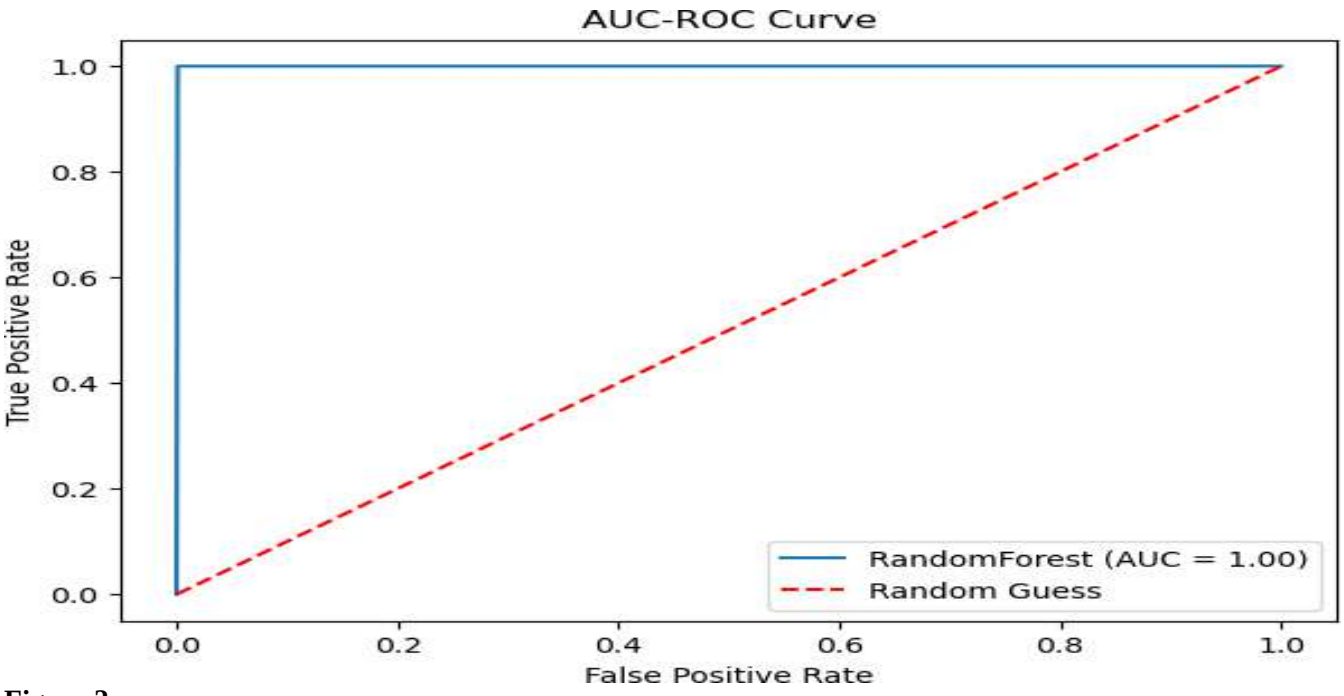

Figure 2
ROC Curve of Random Forest Model predicting MSME Success

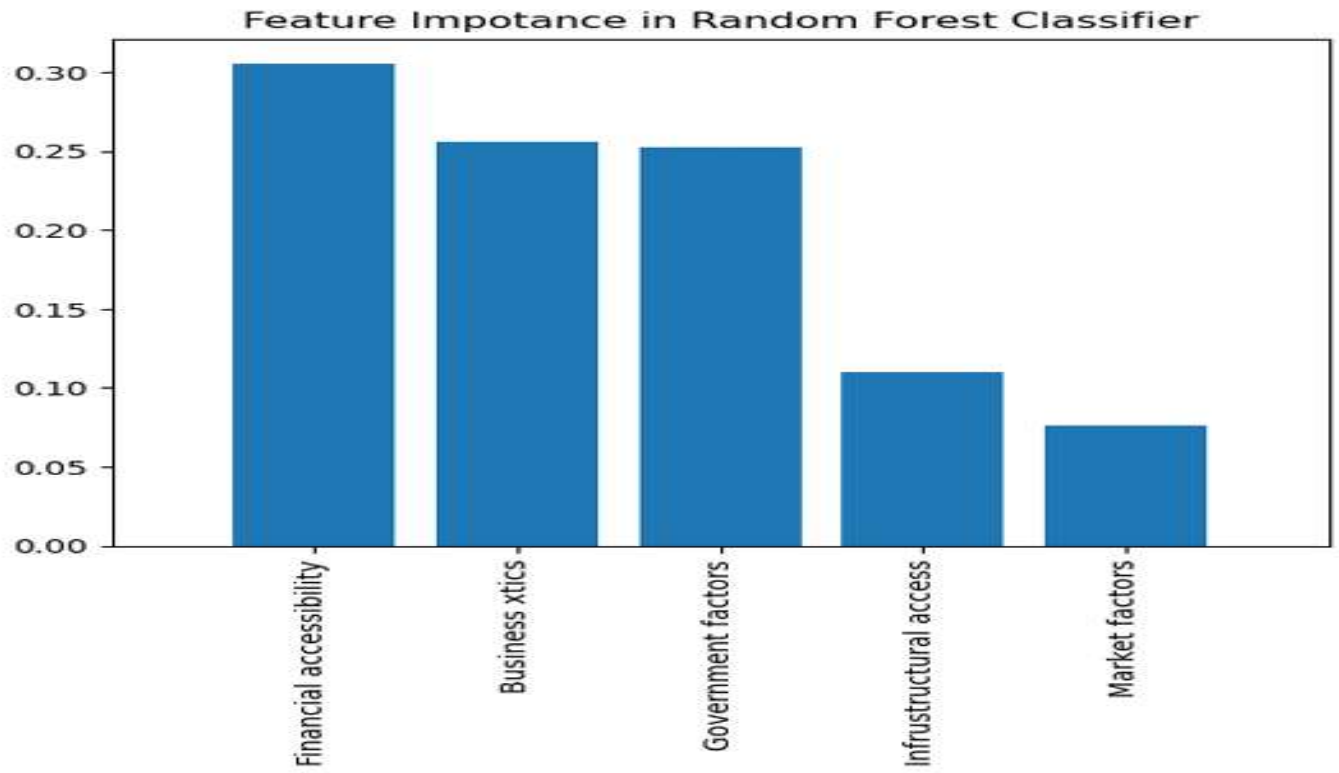

Figure 3
Random Forest Model Importance Scores of Ranked Features

4.2 Logistic Regression Results

Comparatively, the binary logistic regression model was used to confirm the Random Forest results. It accounted for 99.64 percent of the variance in the outcome of MSMEs, and it justified the importance of similar predictors. Logistic regression, although it presupposes linear relations, gave clear coefficients that confirmed the high impact of financial access, business characteristics and government support on the success of business. Being a little less precise than the Random Forest model, it was still interpretable and statistically transparent, which guaranteed the consistency and reliability of the identified success factors in both modeling methods.

```
#Logistic Regression
accuracy_lr = accuracy_score(y_test, lr_model_y_pred)
print(f'Accuracy: {accuracy_lr * 100:.2f}%')
```

```
Accuracy: 99.64%
```

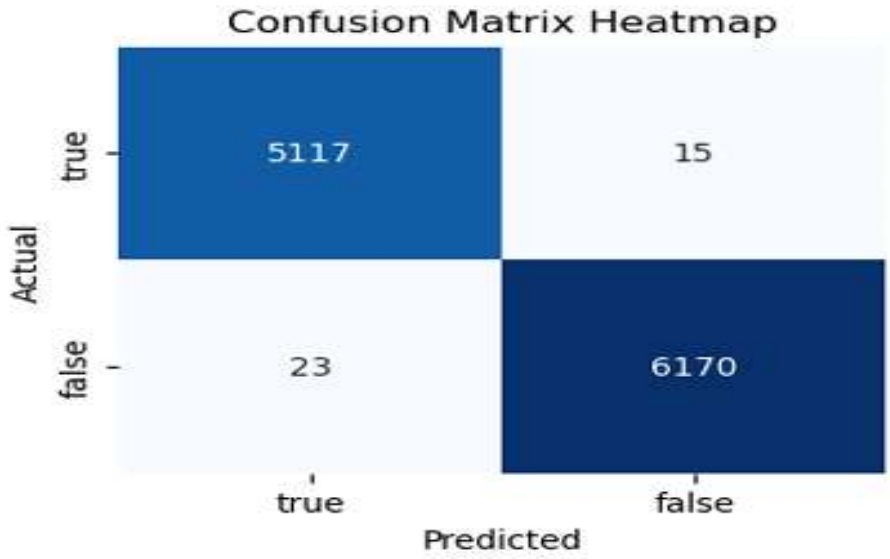

Figure 4
Confusion Matrix of the Logistic Regression Model that Predicts Success of MSME

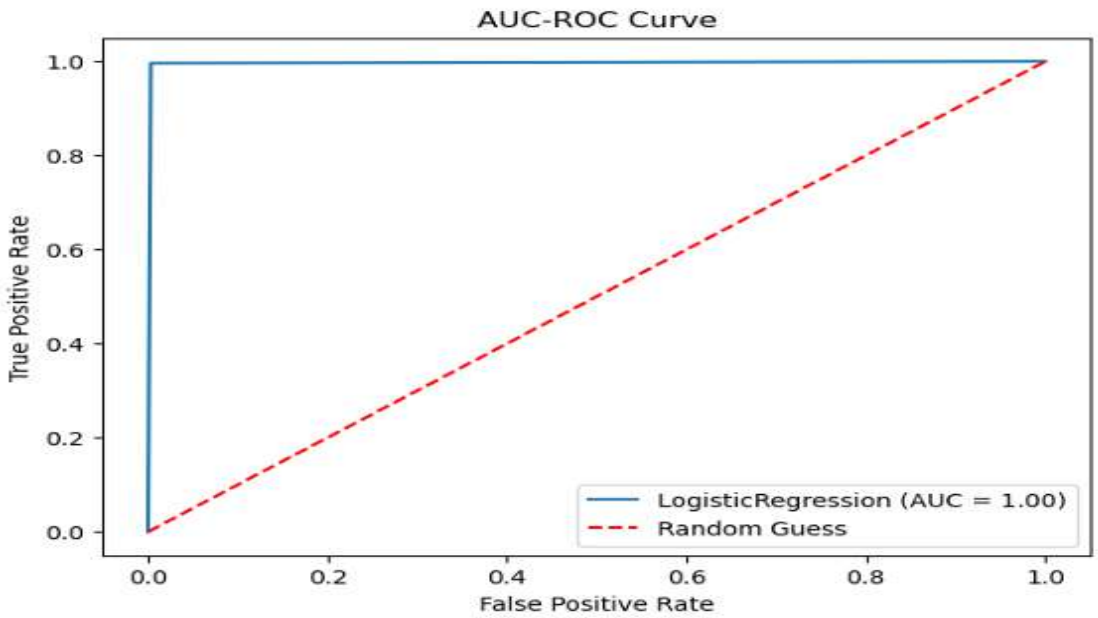

Figure 5
ROC Curve of Logistic Regression Model of MSME Success

4.3 Logistic Regression Analysis

According to the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$ where the coefficients of the above mentioned variables are x_1 =financial access, x_2 = infrastructural access, x_3 = market factors, x_4 = business characteristics and x_5 = government support and $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the coefficients of the respective variables stated above, the study uses logistic regression to generate the coefficients for the above equation which forms the basis of the model in this research.

```
feature_names = ['Financial accessibility', 'Infrustructural access', 'Market
 factors', 'Business xtics', 'Government factors']
coefficients = lr_model.coef_
intercept = lr_model.intercept_
coefs = []
for i in coefficients[0]:
    coefs.append(round(i, 2))

coef_dict = dict(zip(feature_names, coefs))

print("Coefficients with feature names", coef_dict)
print("Intercept:", round(intercept[0], 2))
```

```
Coefficients with feature names {'Financial accessibility': np.float64(6.87),
'Infrustructural access': np.float64(1.3), 'Market factors': np.float64(1.58),
'Business xtics': np.float64(6.67), 'Government factors': np.float64(6.87)}
Intercept: 1.62
```

Based on the above results of the logistic regression, the model which can be used to predict the success of SME is y = 0.16 +0.69x $_1$ + 0.30x $_2$ + 0.16x $_3$ + 0.67x $_4$ + 0.69x $_5$ which predicts the log odds of a business being successful. For normalization purpose, the analysis of logistic regression coefficients is given as probability of the factors which contribute to SME success. Since the coefficient values are {0.16, 0.69, 0.13, 0.16, 0.67, 0.69} where x_1 =financial access, x_2 = infrastructural access, x_3 = market factors, x_4 = business characteristics and x_5 = government support, the coefficients are normalized based on soft-max normalization method as given by

$$x_i = \frac{e^{x_i}}{\sum e^{x_i}}$$

The probability coefficients after normalization are {0.001, 0.353, 0.001, 0.002, 0.289 and 0.353}.With this equation, the model of SME success is transformed into y= $0.001x_0$+ $0.353x_1$+ $0.001x_2$ + $0.002x_3$+ $0.289x_4$+ $0.353x_5$, where + is an AND operator. For this reason, it is observable that, the probability of financial access and government support determine the SME success by 35 percent and 35 percent respectively. It is also revealed that business characteristics have a contribution of about 28 percent to the SME success and that the access to infrastructure and market factors do not have significant contribution to the SME success.

4.4 Discussion

The study effectively addressed a significant gap in the conventional assessment of MSME performance by successfully developing and evaluating a highly accurate predictive model based on the Random Forest Algorithm. This innovation directly counters the limitations of traditional models which often struggle to handle complex, informal and dynamic nature of MSMEs in Kenya leading to ineffective policy-making and hindering sustained growth.

The study's successful application of the Random Forest model with its high accuracy directly aligns and validates the propositions found in the broader academic literature on machine learning. Scholars and development experts widely believe that using predictive models based on machine learning can improve the accuracy of the insights. The literature specifically identifies Random Forests as a suitable choice for analyzing the noisy, incomplete and unstructured data characteristic of MSMEs in developing economies. The predictive model's accuracy serves as a solution to the identified gap. The Random Forest model achieved an accuracy of 99.72 percent in classifying businesses according to the factors that influence their success. This was further reinforced by binary logistic regression results, which indicated that the identified predictors explain 99.64 percent of the variation in the outcome.

The study's simulation findings strongly align with existing knowledge about how MSMEs performance in developing nations. Financial accessibility and government support factors emerges as most statistically significant predictors. This aligns with MSME literature in Kenya and Sub-Saharan Africa where credit access and policy support are repeatedly highlighted as critical drivers (Mwania & Muganda, 2019). While relatively less impactful, infrastructure and market factors still showed significant coefficients, suggesting their supportive role rather than

being sole determinants of business success (Siggelkow & Fernandez, 2024). The findings also confirm that access to credit is vital factor in helping MSMEs succeed, as noted by Beck and Demirgüç-Kunt (2006).

Moreover, the logistic regression model revealed strong evidence of the existence of different determinants of MSME success in line with the previous studies that emphasize access to finance and government facilitation as the main drivers of growth (Mwania & Muganda, 2019). The reliability of these findings is strengthened by the high accuracy of the model (99.66%) and proves the validity of the theoretical framework. Generally, the findings prove that there is a high level of consistency between empirical findings and the conceptual framework, and that extensive support, especially financial and institutional, is critical to the sustainability of MSMEs.

V. CONCLUSION & RECOMMENDATIONS

5.1 Conclusion

The Random Forest Algorithm is effective in predicting MSMEs success, achieving a high accuracy of 99.72 percent in classifying businesses based on critical success factors. The model robustly identified financial access, government support and business characteristics as strong and statistically significant predictors of MSMEs success confirming their pivotal roles. The consistency of these findings further supported by the binary logistic regression, explaining 99.64 percent of the outcome variation validates the model's reliability for informing interventions aimed at enhancing MSMEs success. The predictive model offers prescriptive insights for stakeholders, enabling data-driven decisions that contribute to MSME resilience, growth and sustainable economic development by analyzing key success predictors. The general conclusion is that the incorporation of the advanced data analytics into the MSME development strategies is needed to transition to the proactive support systems.

5.2 Recommendations

Based on the insights derived from the results of this research, a number of strategic suggestions are presented to increase the success and sustainability of MSMEs in Kenya. The study recommends utilization of predictive analytics for decision making. Financial institutions and development partners are strongly encouraged to adopt and leverage predictive models such as the one developed in this study for data-driven decision making in risk assessment, resource allocation and targeted support for MSMEs. The policymakers are advised to focus on the increased availability of affordable financial services and enhancing government support programs that directly address MSMEs, especially those in informal or rural contexts. The financial institutions are also advised to use predictive models, including the Random Forest algorithm, to enhance the credit risk evaluation and find promising businesses that might not have a traditional credit history or collaterals. Finally, future studies must take into account the possibility of using this model in other areas and industries to determine its scalability and improve its predictive power in other economic conditions.

REFERENCES

Barney, J. B. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, *17*(1), 99–120.

Beck, T., & Demirguc-Kunt, A. (2006). Small and medium-size enterprises: Access to finance as a growth constraint. *Journal of Banking & Finance*, *30*(11), 2931–2943.

Berger, A. N., & Udell, G. F. (2006). A more complete conceptual framework for SME finance. *Journal of Banking & Finance, 30*(11), 2945–2966.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792.

De Mel, S., McKenzie, D. J., & Woodruff, C. (2009). Measuring microenterprise profits: Must we ask how the sausage is made? *Journal of Development Economics*, *88*(1), 19–31.

Gichuki, J. A., Njeru, A., & Tirimba, O. I. (2014). Access to credit facilities challenges of micro and small enterprises in Kangemi Harambee Market in Nairobi City County, Kenya. *International Journal of Scientific and Research Publications, 4*(12), 1–25.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

KNBS. (2016). *Basic report of micro, small and medium establishments (MSME) survey*. Nairobi: Kenya National Bureau of Statistics.

KNBS. (2022). *Economic survey 2022*. Nairobi: Kenya National Bureau of Statistics.

Kinyanjui, M. N. (2014). *Women and the informal economy in urban Africa: From the margins to the centre*. Zed Books Ltd.

Mwania, J., & Muganda, R. (2019). The impact of access to finance and government policies on MSME growth in Kenya. *Journal of Small Business Studies, 2*(1), 45–53.

Siggelkow, C., & Fernandez, R. M. (2024). Random forest prediction of SME default with nonfinancial variables: An empirical study of German firms. *Journal of the International Council for Small Business, 5*(2), 129–147.