International Journal of Science and Social Science Research [IJSSSR] ISSN: 2583-7877

Development Of a Poverty Prediction Model Using Geospatial Data in The Oshikoto Region, Namibia

Paulus M Shituna[1]; Nalina Suresh[2]

[1]Department of computing, mathematical & statistical sciences, University of Namibia

[2]Department of computing, mathematical & statistical sciences, University of Namibia

Corresponding Author Email: shitunakadoza@gmail.com

Abstract— This study addresses the challenge of eradicating poverty in developing nations by exploring machine learning (ML) as a tool for efficient poverty prediction. Traditional poverty assessments rely on decennial household surveys, which are resource-intensive and infrequent, especially in African countries. The research leveraged census data from Namibia's Oshikoto Region, training three ML models - Logistic Regression, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LGBM) to classify households as poor or non-poor. The models were trained with 10-fold cross-validation using two feature selection methods: Filter and SHAP.

With the full feature set (350 variables), the study achieved a maximum prediction accuracy of 88.21%. Using the SHAP method, the top 50 features achieved 85.23% accuracy, while the top 20 and 10 features yielded accuracies of 83.67% and 75.79%, respectively. In contrast, the Filter method significantly underperformed, achieving 63.99% accuracy with the top 50 features. Additional metrics, including Area under the curve (AUC), receiver operating characteristic curve (ROC), precision, recall, and Cohen's kappa, confirmed the models' reliability.

The findings demonstrate the feasibility of using ML for accurate poverty prediction with reduced feature sets, highlighting the SHAP method's effectiveness in preserving accuracy. This enables shorter, cost-effective surveys to be conducted more frequently, empowering policymakers and aid organizations to target resources effectively. By focusing on explanatory features, the study provides a scalable framework for addressing poverty in Oshikoto and similar regions, ensuring timely interventions and better resource allocation.

Keywords: Extreme Gradient Boosting Machine (XGBoost), Light Gradient Boosting Machine (LGBM), Logistic Regression, Machine Learning (ML), Namibia

I. INTRODUCTION

I.I. ORIENTATION TO THE STUDY

Poverty is a complex and multifaceted issue that affects millions of people around the world (United Nations Development Programme, 2023). Understanding the spatial distribution of poverty is crucial for designing effective policies and interventions to alleviate poverty and promote equitable development (Sachs, Kroll, Lafortune, Fuller, & Woelm, 2021). In recent years, the use of spatial data and advanced analytical techniques has emerged as a promising approach for predicting poverty and identifying areas of high vulnerability (Naude, 2019).

According to a 2023 report by the Namibia Planning Commission (NPC, 2023) despite relatively good economic and slightly employment growth, poverty continues to disproportionately affect rural households. Nevertheless, the Namibian population remains vulnerable to poverty, a condition further exacerbated by an unemployment rate currently standing at 36.9%. Namibia, classified as a high middle-income country, has some of the highest levels of poverty and inequality. The NPC report indicates that about 28.7% of the population lives in poverty, with 15% being extremely poor. Poverty rates are higher in rural areas (37%) compared to urban areas (15%), and higher among women (32%) than men (26%). Regions with predominantly rural populations, such as Kavango, Zambezi, Oshikoto, Otjozondjupa, Omaheke, Ohangwena, and Kunene, have poverty rates exceeding the national average. In contrast, more urbanized regions like Khomas and Erongo have poverty rates of 10% or less.

The report highlights that the poor are predominantly less educated, with 80% having only primary education or none at all. Just 17% of the poor have completed secondary education, and poverty among graduates is nearly non-existent at less than 1%. This underscores the importance of education in reducing poverty. Poverty rates are highest among old age pensioners (44%) and subsistence farmers (39%), while the working poor account for 16%.

Poverty is also most severe among those with limited access to essential services. It is highest among people whose main source of drinking water is a river, canal, lake, Oshana, dam, well, or public pipe. In terms of sanitation, those without any toilet facilities or who use the bucket system are the poorest. Additionally, those who neither own nor have access to telephones, radios, livestock, or land for crops or grazing are most affected by poverty. Furthermore, individuals living more than a kilometre away from their main sources of drinking water, health facilities, schools, public transport, and local markets are more likely to be poor.

According to the Namibia Statistics Agency (NSA, 2023), Namibia exhibits a high Gini coefficient of 0.597, placing it among the world's most unequal countries. This inequality is more pronounced in urban areas (0.583) compared to rural areas (0.487). Interestingly, a trend of decreasing inequality is observed in rural regions, while urban areas have seen a slight increase. Geographically, the data reveals the highest level of inequality in ǁKharas region (0.634) and the lowest in Ohangwena (0.405). The report further suggests that educational attainment has a minimal impact on overall inequality levels. However, individuals deriving their primary income from business (0.656) and salaries/wages (0.568) demonstrate the highest levels of income inequality.

Highlighting the value of spatial data, Sliuzas (2016) emphasizes its role in poverty research. Spatial data, encompassing Geographic Information System (GIS) data, locational information, maps, satellite imagery, and socio-economic indicators, offers crucial insights into the geographical patterns and factors associated with poverty. By analysing the relationships between these diverse spatial variables and poverty indicators, researchers gain a deeper understanding of the root causes of poverty, enabling the development of targeted strategies for its alleviation. This spatial understanding is particularly valuable for policymakers, government agencies, and poverty-reduction organizations. By leveraging spatial data analysis, these entities can predict areas with high poverty concentrations, identify spatial patterns of poverty, and uncover the underlying factors contributing to it.

The use of spatial data for poverty prediction offers several advantages. First, it allows for a more granular analysis by considering the spatial context of poverty. This enables the identification of specific areas or communities that are most at risk, facilitating the allocation of resources and interventions where they are needed the most. Second, spatial data can capture the influence of local environmental and socio-economic factors on poverty, providing valuable insights into the underlying mechanisms and drivers. Third, the integration of spatial data with predictive models enhances the accuracy and robustness of poverty predictions, enabling evidence-based decision-making.

However, predicting poverty using spatial data also poses challenges. Data availability and quality, spatial scale, and model selection are important considerations that need to be addressed. Additionally, the dynamic nature of poverty requires the incorporation of temporal aspects and the ability to capture changes over time.

In this context, the study aimed to explore the potential of census data and predictive modelling techniques for poverty prediction in Oshikoto Region. By analysing a range of spatial variables, including demographic, socio-economic, and environmental factors, the study sought to develop a predictive model that can accurately estimate poverty levels and identify areas of high vulnerability. The findings of this research can inform policy-makers, organizations, and stakeholders involved in poverty alleviation efforts, enabling more targeted and effective interventions.

I.II. PROBLEM STATEMENT

The Namibia Statistics Agency (NSA) maintains comprehensive datasets, including the Namibia Household Income and Expenditure Survey (NHIES) and the Namibia Population and Housing Census. However, despite possessing these data resources, the agency currently does not have an effective forecasting model to predict and map poverty levels, severity, and trends for informed developmental planning and intervention. As noted by the NSA's Chief Executive Officer, Shimuafeni (2021), 'the agency presently doesn’t have a poverty prediction system in place, although there are plans to develop one in the future.' The absence of this predictive system obstructs the realization of the government's National Development Plan (NPD6) agenda aimed at improving living standards in various villages within Oshikoto region by 2030, despite the NSA possessing the necessary datasets. Consequently, this deficiency contributes to regional disparities, hindering economic development, infrastructure, environmental enhancement, and the overall improvement of living conditions. Conventional methods have measured poverty levels using survey data. However, research has shown that door-to-door surveys are expensive and time-consuming (Jean, Burke, Ermon, Cantú, & Gobron, Combining satellite imagery and machine learning to predict poverty, 2016). Census surveys in Namibia are undertaken every decade, contingent upon the availability of funding from local and international

development agencies. This indicates that poverty estimates for the population are unavailable between survey years. This creates a problem of policy design and decision-making because old data are used to design development programmes, leading to inaccurate results (Jerven, 2013).

Therefore, there is an urgent need for a robust, geospatially-enabled poverty prediction model that can leverage existing datasets to generate timely, spatially detailed, and accurate poverty estimates. Such a system would enable policymakers and development agencies to identify high-poverty areas, design targeted interventions, allocate resources more efficiently, and accelerate progress towards Namibia's development goals, particularly in addressing rural poverty in the Oshikoto region.

I.III. OBJECTIVES OF THE STUDY

Main objective(s)

To develop a poverty prediction model using the geospatial indicators and estimate the poverty level of the households in Oshikoto Region.

Sub-Objectives are to:

a) To extract the appropriate 2023 Oshikoto Region poverty census dataset for the development of a poverty prediction model.

b) To apply machine learning techniques to develop a prediction model to accurately predict poverty in the Oshikoto Region.

c) To test and evaluate the model performance for optimal prediction.

I.IV. SIGNIFICANCE OF THE STUDY

The significance of this study lies in its ability to leverage census data and predictive modelling techniques to improve poverty prediction, inform policy-making, optimize resource allocation, promote spatial equity, and contribute to the overall effectiveness and impact of poverty reduction efforts. The present study seeks to evaluate poverty levels via machine learning classification techniques rather than survey-to-survey imputation. This experiment utilises freely available census survey data from the Namibia Statistics Agency, derived from decennial geographical surveys, instead of employing the more costly data processing approaches often utilised in machine learning.

The machine learning frameworks, including PyCaret and Scikit-learn, are readily available, open-source tools that can operate on minimal computing resources. This study further investigates the evaluation of poverty levels by employing a selected set of characteristics to determine if comparable prediction accuracy can be attained with fewer variables than when utilising the complete array of variables available from integrated home survey data.

The application of machine learning for poverty classification can facilitate the cost-effective and precise monitoring of poverty trends, enabling policymakers to implement poverty reduction measures swiftly and enhance the identification and targeting of the most vulnerable villages within the region.

I.V. LIMITATIONS OF THE STUDY

This study exclusively utilized data from the 2023 Population and Housing Census (PHC) for the Oshikoto region. Census data from previous years were excluded because they were manually collected, potentially containing inconsistencies, errors, and missing values. In contrast, the 2023 census data, being fully digitalized, provided greater accuracy and consistency. Additionally, due to computational limitations and constrained resources, the modelling process employed a ten-fold cross-validation method, partitioning the dataset into training and testing subsets.

I.VI. DELIMITATION OF THE STUDY

The geographical scope of this study was confined to the Oshikoto region. This delimitation restricts the generalizability of the findings to other regions with potentially different socioeconomic characteristics and poverty indicators. Consequently, the models' accuracy might be optimized for the Oshikoto region and may require further evaluation or adaptation for application in other contexts.

I.VII. CHAPTER SUMMARY

This chapter introduced the study by providing an orientation to the research context and highlighting the pressing challenge of poverty in Namibia, particularly within the Oshikoto region. The problem statement identified the absence of a reliable and geospatially enabled poverty prediction model, which limits timely, evidence-based decision-making for targeted interventions. The study's objectives were outlined, aiming to develop a robust model to predict and map poverty levels, severity, and trends using existing datasets. The significance of the research was emphasized in its potential to inform policy, improve resource allocation, and contribute to Namibia's developmental goals. The chapter also acknowledged the limitations and delimitations that frame the scope and boundaries of the investigation. Collectively, these elements establish the foundation and rationale for the subsequent chapters of this thesis.

II. LITERATURE REVIEW

II.I. INTRODUCTION

The chapter examines methodologies for poverty forecasting utilising machine learning techniques. It accomplishes this by examining the used data types, the employed machine learning tools and methodologies, and the performance outcomes of the models in the investigations. The utilised data types comprise geospatial data obtained from completed questionnaires, cell phone records, and satellite imagery. Predicting poverty has gained significant attention in the field of poverty analysis and has been the subject of numerous studies (Sultana, Arif, & Hasan, 2023).

II.II. SUMMARY OF MACHINE LEARNING (ML)

Advancements in computer architecture and memory organization since the 1990s have significantly enhanced microprocessor performance, facilitating the growth of machine learning applications (mdpi, 2021). ML aims to analyse data to discover patterns, which can then be applied in different contexts to generate similar results with minimal human input (Géron, 2019). The following are different type of machine learning.

II.II.I. REINFORCEMENT MACHINE LEARNING

Reinforcement learning (RL) is a machine learning paradigm that emphasises decision-making by autonomous agents.

This approach is particularly effective for solving problems that involve sequential decision-making and interaction (Sutton & Barto, 2018). Instead of learning from a static dataset, the agent explores different strategies and continuously updates its knowledge based on the outcomes of its actions, aiming to maximize cumulative rewards over time (Jacob & Kavlakoglu, 2024).

Reinforcement learning fundamentally involves the interaction of an actor, an environment, and a goal. Literature extensively articulates this link using the Markov decision process (MDP). In the context of a Markov decision process, the agent serves as the decision-making entity within the reinforcement learning framework. It engages with the environment by executing actions to optimise cumulative rewards. The state (St) denotes the present status of the environment. It furnishes all required information for the agent to determine an action. Action (At) refers to the response of the agent to a specific situation. The agent chooses an action from a range of potential actions to affect the surroundings. Reward (Rt) is a quantitative measure that assesses the efficacy of the agent's action. A positive reinforcement strengthens desirable behaviour, whereas a negative reinforcement deters undesirable choices. State Transition (S(t+1)) Subsequent to an action, the environment undergoes a transition to a new state. This is denoted as S(t+1)R(t+1), indicating the subsequent state following the execution of action At. Reward Update (R(t+1)) Upon state transition, the environment delivers a new reward, R(t+1). The agent utilises this feedback to modify subsequent decisions. A typical reinforcement learning process can be described as depicted in Figure 1 below.

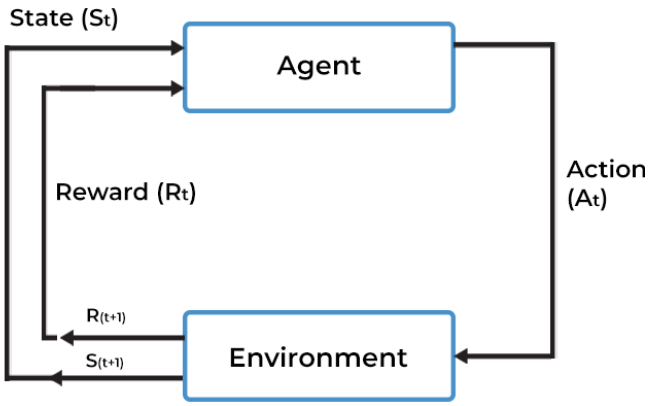


Figure 1.A Pictorial Representation of the Reinforcement Learning Model

Figure Source: (Springer, 2023)

The state space encompasses all information sent by the state of the environment. Action space denotes all possible actions the agent may take within a state (Brandon & Alexander, 2020).

II.II.II. SUPERVISED MACHINE LEARNING

Supervised learning is a category of machine learning in which a model is trained using labelled data, signifying that each input is associated with the corresponding output (Raschka & Mirjalili, 2019). The model acquires knowledge by contrasting its predictions with the actual responses presented in the training data. Another subset, regression analysis, aims to predict continuous outcomes (Raschka, Python Machine Learning (1st ed.), 2020).

The objective of supervised learning is to develop a function that predicts a target variable utilising pre-labelled input data. This entails developing a model that analyses fresh data to forecast the target variable. According to the Korivi (2016), illustrates supervised learning in classification with an example: an algorithm can be trained to identify an animal by using a dataset of pre-labelled images of various species (Korivi, 2016).

A supervised learning algorithm comprises input features and their associated output labels. The procedure operates as follow:

a) Training: The model is supplied with a training dataset comprising input data (features) and associated output data (labels or target variables).
b) Learning: The algorithm analyses the training data, discerning the correlations between the input attributes and the output labels. This is accomplished by modifying the model's parameters to reduce the disparity between its predictions and the actual labels.

Subsequent to training, the model undergoes evaluation with a test dataset to assess its accuracy and performance. The model's performance is enhanced by fine-tuning parameters and employing methods such as cross-validation to equilibrate bias and variance. This guarantees the model generalises effectively to novel, unobserved data.

II.II.III. UNSUPERVISED MACHINE LEARNING

Unsupervised learning algorithms are designed to identify patterns and relationships within data without any prior understanding of the material's significance (Trevor, Robert, & Jerome, 2017). Unsupervised machine learning algorithms identify concealed patterns in data autonomously, without human interaction, meaning no output is provided to the model. The training model possesses solely input parameter values and autonomously identifies groups or patterns. The input data in these methodologies is unannotated and devoid of labels. Unsupervised machine learning algorithms discern patterns and categorise data into clusters according to feature similarity. However, they face challenges such as complexity, high computational resource demands, and difficulties in feature selection (Williams, Zander, & Armitage, 2006).

II.II.IV. MACHINE LEARNING DEVELOPMENT FRAMEWORK

The framework for Raschka (2020) describes a standard procedure for developing a machine learning system, with four primary phases. The initial phase is data pre-processing, which readies the data for utilisation by any machine learning algorithm to attain improved outcomes. Figure 2 in the linked section demonstrates that numerous ML algorithms have been created to tackle diverse

challenges. During the training phase of machine learning development, an algorithm is instructed using pertinent input data to create a prediction model designed for a certain purpose. This phase entails supplying the algorithm with data, facilitating its identification of patterns and relationships, which in turn permits precise predictions or judgements when confronted with novel, unknown data. Subsequently, model evaluation occurs, which appraises the model's efficacy utilising established measures. Upon satisfying the performance criteria, the model may be deployed to generate predictions on novel or unobserved data. Raschka (2020)'s workflow schematic is presented in Figure 2 below.

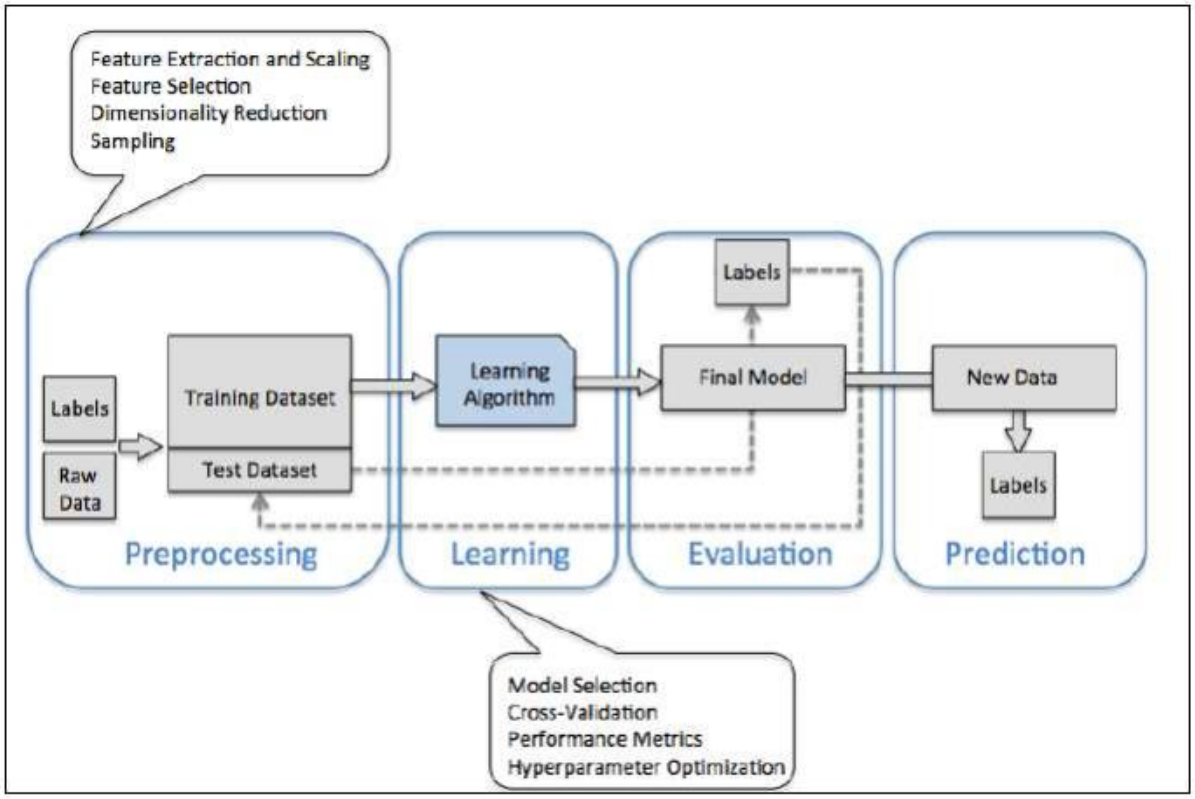


Figure 2.Predictive model Schema

Figure Source: (Raschka, Python Machine Learning (1st ed.), 2020)

II.II.V. MACHINE LEARNING TYPES SUMMARY

Understanding the different types of machine learning provides a foundation for selecting algorithms that align with the goals and data characteristics of this research. Since the aim is to develop a predictive model for poverty in the Oshikoto region using socio-economic and geospatial survey data, supervised learning is most appropriate. This approach allows the model to learn from labelled training data, where poverty status is already known, and then make predictions for new, unseen cases. Within supervised learning, classification techniques are particularly relevant as they enable the categorisation of households into poverty levels, thereby supporting evidence-based targeting of interventions.

II.III. AN EXAMINATION OF RESEARCH UTILISING MACHINE LEARNING FOR POVERTY PREDICTION

Machine learning (ML) techniques have been successfully applied across various fields. Recently, they have garnered interest in predicting economic well-being and enhancing poverty alleviation efforts (Mullainathan & Spiess, 2017). These studies have employed various data sources such as, mobile cell phone data, satellite imagery, and survey data. The following segment examines research from the past decade that utilizes machine learning to assess poverty levels within Africa and beyond.

II.III.I. MACHINE LEARNING CLASSIFICATION USING SURVEY DATA

Machine learning algorithms have been utilised independently on survey data, as well as in conjunction with other data types for the purposes of classifying poverty, validating, or supplementing findings. Below are several studies that employed machine learning classification techniques on survey data to forecast poverty.

In this work, Fitzpatrick, Bull, & Dupriez, (2018) juxtaposed the outcomes of optimised linear regression with ten contemporary classification techniques, encompassing sophisticated decision tree methodologies, genetic algorithms, and deep learning approaches (Fitzpatrick, Bull, & Dupriez, 2018). Consequently, they could evaluate the trade-offs between the computational complexity of the methods and the performance of the model. The employed categorisation algorithms were entirely open source and utilised on survey data from Malawi and Indonesia. Survey results were transformed into attributes utilised as predictors to classify households as "poor" or "non-poor." The findings revealed no significant disparity in the prediction efficacy of linear regression compared to more sophisticated machine learning techniques. Indeed, in all instances, linear regression outperformed the more sophisticated techniques, ranking among the top three predictors. The forecast accuracy values were from 73% to 87% for Malawi and from 88% to 91% for Indonesia prior to model fine-tuning, and from 86% to 87% for Malawi and from 81% to 85% for Indonesia subsequently. In their uncomplicated data set comprising 10 features, model accuracy varied between 73% and 77%. Significantly, their top 10 traits were obtained by a stepwise methodology similar to that employed in regression procedures for selecting primary predictors.

In 2019, æbø, Ø. A., Wiig, H., & Hatlebakk, M conducted a study titled "Challenges in Predicting Poverty Trends Using Survey-to-Survey Imputation: Experiences from Malawi." The researchers employed a statistical model that combined comprehensive consumption surveys with lighter annual surveys to monitor poverty trends over a decade in Malawi. Their analysis revealed that incorporating household demographic composition into the model enhanced the accuracy of poverty predictions. However, the study faced significant challenges related to the comparability between different surveys. Variations in survey design and data collection methods posed difficulties in achieving consistent and reliable poverty estimates over time. The findings indicated that the approach attained a low error rate of 19.3% in distinguishing between impoverished and non-impoverished.

In 2023, Fobi, S et al. authored the study "Poverty Rate Prediction Using Multi-Modal Survey and Earth Observation Data." This research integrated traditional household demographic and living standards survey data with features derived from satellite imagery to predict regional poverty rates. The approach resulted in a reduction of prediction errors from 4.09% to 3.71%, indicating an improvement in accuracy by incorporating multi-modal data sources. Despite these advancements, the study encountered challenges in integrating diverse data types, particularly in aligning satellite imagery with survey data. Ensuring the quality and resolution of satellite data to match the granularity required for accurate poverty prediction was also a significant concern.

McBride and Nichols (2018) utilised survey data to forecast poverty through cross-validation and stochastic ensemble techniques, enhancing the predictive capacity of proxy means test (PMT) instruments utilised by the United States Agency for International Development (USAID). The study determined that employing a stochastic ensemble methodology might expedite feature selection and enhance the evaluation of various machine learning models' performance. The ensemble approach demonstrated a modest improvement in performance compared to alternative methods. The predicted accuracy values ranged from 87% for Malawi to 55% for Bolivia.

The difficulties associated with obtaining private data and the computational demands for high-resolution images applicable to other data types do not pertain to the use of survey data.

II.III.II. MACHINE LEARNING CLASSIFICATION USING SATELLITE IMAGERY AND GEOSPATIAL DATA

In 2016, Jean et al. introduced a novel approach that leveraged high-resolution satellite imagery and machine learning to predict economic well-being in five African countries: Nigeria, Tanzania, Uganda, Malawi, and Rwanda. Their method involved training a convolutional neural network (CNN) to identify visual features in satellite images that correlated with economic indicators, ultimately explaining up to 75% of the variation in local economic outcomes. This breakthrough provided an inexpensive and scalable solution for economic assessment, particularly in regions with limited access to mobile phone data. By utilizing publicly available imagery, their approach allowed for broader geographic coverage compared to prior methods that relied on mobile metadata.

Building upon this foundation, Yeh et al. (2020) further advanced the use of machine learning and satellite imagery for economic prediction. Their study expanded the application of deep learning across nearly 20,000 African villages, significantly increasing the dataset size and geographic diversity compared to Jean et al.’s work. Notably, their models achieved 70% accuracy in predicting village wealth, even in countries where the model had not been trained, demonstrating improved generalizability. This advancement addressed one of the key limitations of earlier studies - scalability across different regions - making the method

more robust for cross-country economic assessments. By refining model architectures and leveraging larger datasets, Yeh et al. improved the predictive power and adaptability of satellite-based economic analysis.

Recent studies have continued to refine and expand the use of satellite imagery and machine learning for poverty estimation.

In 2023 study by Fobi et al. introduced a method that combines household demographic surveys with features derived from freely available Sentinel-2 satellite imagery to predict regional poverty rates. Their approach utilized visual features from 10-meter resolution images, integrated with survey data, to estimate household poverty levels. The inclusion of satellite-derived features reduced the mean error in poverty rate estimates from 4.09% to 3.88%. Additionally, they proposed a survey variable selection method to identify the most relevant questions that complement the visual features, further enhancing prediction accuracy. This advancement underscores the ongoing evolution of combining diverse data sources to improve the precision and applicability of poverty mapping models across different contexts.

Although machine learning applied to satellite imagery has attained differing degrees of success in forecasting poverty for economic clusters, it fails to produce results at the provincial or regional level. Moreover, acquiring spatial-temporal satellite imagery for input into the machine learning method poses significant challenges for low-resource nations such as Namibia, owing to the considerable computer power necessary for processing high-resolution satellite photos.

A 2021 study in Myanmar by Liu and Sun, explored the use of OSM-derived features, such as road density and environmental pollution indices, to predict poverty levels. The study found a strong correlation between increased road density, higher pollution levels, and poverty rates. Multiple machine learning models were tested, with the Random Forest algorithm demonstrating the highest predictive accuracy, achieving an R^2 score of 0.79. This study demonstrated the effectiveness of geospatial data in poverty estimation (Liu, et al., 2021). Nonetheless, a significant obstacle of this method is the requirement for subject expertise to filter and annotate geospatial pictures for poverty-related characteristics.

II.III.III. MACHINE LEARNING CLASSIFICATION APPROACHES USING MOBILE PHONE DATA

Toole, Eagle, and Bhattacharya (2017) used an advanced methodology that integrated mobile phone usage statistics with geospatial and demographic characteristics to categorise poverty levels across various locations. The authors initially pre-processed extensive mobile phone data to extract variables including call volume, frequency, mobility patterns, and social network metrics. These attributes were subsequently combined with geospatial variables such as population density, proximity to urban centres, and service accessibility, together with demographic data obtained from census records. A crucial aspect of their methodology included feature engineering and normalisation to alleviate the impact of temporal variability intrinsic to mobile usage patterns, such as variations in call activity between weekdays and weekends or across different seasons. The model training utilised advanced supervised learning methods, achieving an overall classification accuracy of roughly 75%. The study indicated that the model attained superior performance metrics in metropolitan areas, characterised by high mobile phone usage and strong data quality, frequently exceeding 80% accuracy. In contrast, in rural regions characterised by limited mobile activity and inconsistent data gathering, the accuracy decreased to approximately 65–70%. These inconsistencies underscore the difficulties encountered by academics in achieving data representativeness across varied geographic and socioeconomic contexts. The prediction results of the streamlined models were thereafter compared to this index to ascertain whether an individual classified as poor or non-poor.

In 2017, Blondel et al. conducted a study on poverty in Bangladesh that sought to integrate mobile phone metadata with traditional survey data to generate high-resolution poverty maps. Their approach utilized call detail records and other behavioural indicators to predict socioeconomic status, reporting an R^2 score of around 0.65 in validating their model against ground-truth survey data. The study highlighted several challenges, notably the difficulty of reconciling sparse mobile data in rural regions with comprehensive socioeconomic indicators, as well as issues of data privacy and representativeness that could potentially skew the model's accuracy. This work contributed to the growing literature on using unconventional data sources, such as mobile phone data, for poverty estimation in developing countries, and it underscored the importance of methodological rigor and ethical considerations in leveraging digital data for socioeconomic research.

Utilising mobile phone data for poverty prediction provides individual-level insights and serves as a feasible alternative in the absence of survey data. However, this technique neglects certain population segments, especially the very destitute, as not all homes have cell phones. Moreover, Blondel et al. (2017) emphasised the reluctance of mobile phone carriers to reveal data, citing concerns related to business and privacy. This highlights a considerable issue in relying on the availability of corporate data for poverty monitoring.

II.IV. AN INVESTIGATION OF MACHINE LEARNING TECHNIQUES FOR POVERTY PREDICTION

Research in poverty prediction has spanned a variety of data sources and machine learning methodologies, demonstrating both promising performance metrics and distinct challenges.

In the realm of survey data, Fitzpatrick, Bull, and Dupriez (2018) compared optimized linear regression with ten modern classification techniques—including decision trees, genetic algorithms, and deep learning—using datasets from Malawi and Indonesia. Their study transformed survey responses into predictor variables to classify households as "poor" or "non-poor," reporting forecast accuracies ranging from 73% to 87% for Malawi and 88% to 91% for Indonesia prior to fine-tuning, with slight adjustments post-tuning. Notably, their analysis revealed that the relatively simple linear regression model performed on par with, or even outperformed, more complex techniques—a finding that underscores the importance of balancing computational complexity with model efficacy. Further advancing survey-based approaches, Sæbø, Wiig, and Hatlebakk (2019) employed a survey-to-survey imputation method to monitor poverty trends in Malawi over a decade. By integrating comprehensive consumption surveys with lighter annual surveys and incorporating household demographic composition, they achieved a low error rate of 19.3% in distinguishing between impoverished and non-impoverished households. However, their work also highlighted significant challenges stemming from variations in survey design and data collection methods, which complicated efforts to maintain consistent and reliable poverty estimates over time. Complementing these studies, Fobi et al. (2023) introduced a multi-modal approach that fused traditional household demographic surveys with Sentinel-2 satellite imagery. Their innovative integration reduced prediction errors from 4.09% to 3.71%, although the study faced hurdles in aligning disparate data types - particularly reconciling the spatial resolution of satellite imagery with survey data granularity. McBride and Nichols (2018) further contributed by applying cross-validation and stochastic ensemble techniques to refine the predictive capacity of USAID's proxy means test instruments, achieving accuracy values that varied significantly across regions - 87% for Malawi contrasted with 55% for Bolivia - thereby emphasizing the role of local data characteristics in model performance.

Machine learning applications using satellite imagery and geospatial data have also gained traction. Jean et al. (2016) pioneered this approach by training convolutional neural networks (cnns) on high resolution daytime and nighttime satellite images to extract features indicative of economic well-being, ultimately explaining up to 75% of the variation in local economic outcomes. Yeh et al. (2020) expanded on this work by applying deep learning models across nearly 20,000 African villages, achieving an impressive 70% accuracy in predicting village wealth even in unseen regions, thus demonstrating enhanced model generalizability and scalability. Additionally, Liu and Sun (2021) explored the use of OpenStreetMap (OSM)-derived features such as road density and environmental pollution indices to predict poverty levels in Myanmar, with the Random Forest algorithm achieving an R^2 score of 0.79. Their study, while affirming the effectiveness of geospatial data, also encountered challenges related to the need for expert annotation to accurately filter and interpret poverty-related characteristics from geospatial imagery.

Mobile phone data offers another promising avenue for poverty prediction. Toole, Eagle, and Bhattacharya (2017) implemented an advanced methodology that combined mobile phone usage statistics such as call volume, frequency, mobility patterns, and social network metrics with geospatial and demographic data. Their supervised learning models attained an overall classification accuracy of approximately 75%, with performance exceeding 80% in urban areas due to higher data density, but dropping to around 65–70% in rural regions where data were sparse and inconsistent. Complementing this approach, Blondel et al. (2017) investigated poverty in Bangladesh by merging mobile phone metadata with traditional survey data to generate high-resolution poverty maps, reporting an R^2 score of about 0.65. Their work underscored challenges such as reconciling sparse mobile data in rural settings with comprehensive socioeconomic indicators and addressing issues of data privacy and representativeness. Collectively, these studies illuminate the multifaceted nature of poverty prediction using machine learning, where the choice of data source from survey responses to satellite imagery and mobile phone records profoundly influences model performance and underscores the inherent trade-offs between accuracy, scalability, and data quality.

II.V. MACHINE LEARNING METHODOLOGY SUMMARY

In summary, the methodology selection for this study was guided by a balance between predictive performance, computational feasibility, and contextual applicability to the Oshikoto region. While complex deep learning approaches offer high accuracy for large, proprietary datasets such as satellite imagery or mobile phone records, their resource demands and data accessibility constraints make them impractical in this setting. Instead, the chosen classifiers - Logistic Regression, Extreme Gradient Boosting Machine (ExGBM), and Light Gradient Boosting Machine (LGBM) - were selected for their demonstrated success in prior poverty prediction studies, their ability to handle structured survey and geospatial data, and their suitability for

implementation in environments with limited computational resources. This alignment ensures that the resulting poverty prediction model would be both technically robust and practically deployable for informing targeted interventions in the Oshikoto region.

III. RESEARCH METHODOLOGY

III.I. INTRODUCTION

This chapter outlines the machine learning (ML) tools and libraries utilized, along with the dataset and its collection process. It describes the research design and methodology employed to achieve the study's objectives.

A robust methodological framework was critical for effectively analysing and predicting poverty levels. The study leveraged established Python-based ML libraries, chosen for their reliability, extensive documentation, and wide adoption within the data science community. These tools facilitated the application of state-of-the-art predictive modelling techniques. The key libraries included; NumPy by (Mohammed & Al-Rawi, 2023) , Jupyter Notebook (Team, 2015) , SciPy, by (Mohammed & Al-Rawi, 2023), Pandas, by (Reback, McKinney, & brockmendel, 2021), Matplotlib, (Caswell, Droettboom, & Lee, 2023), Scikit-learn, by (Buitinck, Louppe, Blondel, & Pedregosa, 2023), PyCaret by (Ali, Jabeen, & Khan, 2022), SHAP (SHapley Additive exPlanation) by (Lundberg, Erion, & Lee, 2020), Keras, by (Gulli, Pal, & Kapoor, 2021), Statsmodels, by (Perktold, Seabold, Taylor, & Brodbeck, 2023) and Seaborn, by (Waskom, 2022). These tools collectively supported data pre-processing, feature selection, model training, and evaluation, ensuring a comprehensive approach to poverty prediction.

III.I.I. RESEARCH DESIGN

The study used the quantitative research approach for data collection, preparation, model building, and results evaluation. Given the focus on predictive analytics and machine learning, the research adopted Raschka's (2019) machine learning workflow methodology. This methodology effectively supported the research objectives by providing a structured framework for developing and deploying machine learning models. The study was structured into three distinct phases, outlined in details in the following phases. The following fig. 1 shows the Raschka's methodology phases for using machine learning in predictive modelling

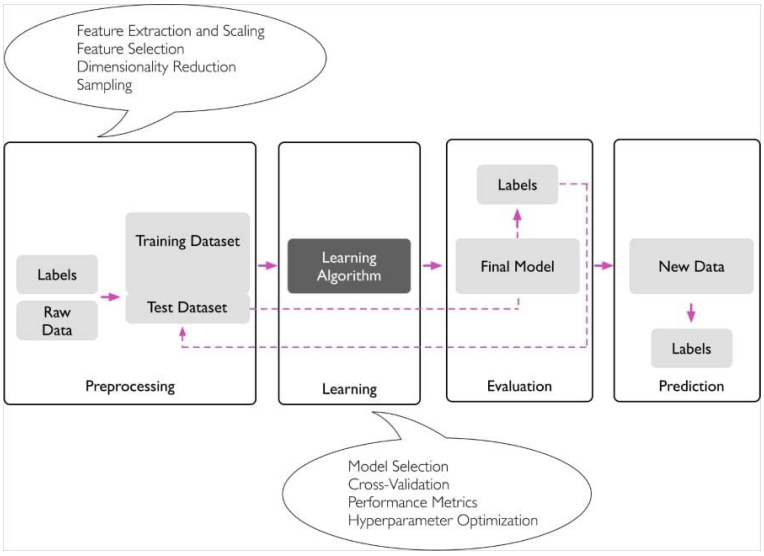


Figure 1. Raschka's methodology phases for using machine learning in predictive modelling

Phase 1: Data Pre-Processing

The initial data pre-processing phase is critical for ensuring data quality. Fitzpatrick et al. (2018) emphasized the importance of this step in machine learning. The dataset for this study, sourced from the NSA portal, underwent systematic pre-processing. Data was first extracted using Stata software and converted into CSV format via the Pandas read_stata() function. Missing values, encoded as blanks or NaNs, were handled using the Pandas drop method. Features were then derived from individual-level survey data to enrich the household dataset, such as calculating the number of children or males per household. Categorical features were transformed using one-hot encoding with the Pandas get_dummies() function, mitigating multicollinearity by dropping the

first category of each feature. Unnecessary features, such as those with one unique value or duplicates, were dropped to improve efficiency.

Exploratory Data Analysis (EDA) played a crucial role in understanding relationships between features and the poverty target variable. Distributions of features were examined, and uninformative features were removed to retain only predictive variables. Feature selection methods, including the filter method and SHAP (SHapley Additive exPlanations), were applied to reduce the dataset's dimensionality. The filter method ranked features based on statistical metrics, while SHAP utilized Shapley values to assess feature importance. This approach enabled the study to evaluate prediction accuracy using both the full dataset (350features) and subsets ranging from 20 to 450 features.

Phase 2: Learning, the focus shifted to training and selecting a predictive model.

The dataset was split into training and testing sets in an 80:20 ratio. PyCaret, an open-source ML framework, was employed to train classification algorithms and evaluate their performance. Logistic Regression (LR), Extra Gradient Boosting Model (ExGBM), and Light Gradient Boosting Model (LGBM) were pre-selected based on their effectiveness in poverty classification. To prevent overfitting, a 10-fold stratified cross-validation approach was adopted, where the dataset was divided into folds, with models trained and tested iteratively across these folds. Performance metrics, including accuracy, recall, precision, F1-score, and AUC, were used to assess and compare model performance. PyCaret's hyperparameter optimization functionality, tune_model(), was used to fine-tune model parameters for optimal performance.

Phase 3: Model evaluation and data prediction

In this phase, the effectiveness of models was assessed using the aforementioned metrics. Predictions were made on the reserved test dataset as a proxy for unseen data. The study evaluated the models' performance across various feature subsets to identify the optimal configuration for accurately predicting poor and non-poor households. This phase aligned with Raschka's (2019) workflow, emphasizing robust model evaluation and reliable predictions for future applications.

III.I.II. DATASET COLLECTION

The study utilized data from Namibian Statistics Agency (NSA)'s Population and Housing Census conducted in 2023-2024, which was collected through the Agency's data portal with ethical clearance certificate from the University of Namibia. This portal is publicly accessible and the data is available under a Creative Commons license.

The dataset comprised two files: one containing household data and the other containing spatial data from Oshikoto Region. The household dataset included population information on 7000 households, classified as either "poor" if their per capita consumption was less than or equal to the poverty line of USD 5.50 per person per day, or "non-poor" if consumption was above the poverty line (Namibian-government, 2023). These households encompassed 56,211 individuals, whose data were recorded in the households file. The household data file featured 346 attributes, while the spatial data table contained 42 attributes.

III.I.III. DATASET ANALYSIS

The dataset underwent initial cleaning and subsequent analysis using Python libraries such as pandas, NumPy, and Matplotlib. Spatial clustering algorithms were employed to visualize and interpret the results.

IV. FINDINGS PRESENTATION

IV.I. INTRODUCTION

This section unveils the outcomes derived from executing the procedures delineated in Chapter 3, aimed at fulfilling the study's objective by addressing its research objectives regarding the feasibility of predicting poverty utilizing readily available machine learning (ML) tools applied to survey data and the potential for predicting poverty with a reduced set of features while maintaining satisfactory outcomes. Presented herein are the findings from data pre-processing, exploration, feature selection, as well as model training and evaluation, with detailed discussions on the results achieved at each phase. Ultimately, the chapter circles back to the initial research objectives, leveraging the obtained results to ascertain their resolutions.

IV.II. FEATURE SELECTION RESULTS

Applying the two feature selection techniques yielded two distinct rankings of the 350features. The following are the top 20 features identified by the Filter and SHAP methods.

IV.II.I. TOP FEATURES SELECTED BY THE FILTER METHOD

When the dataset was processed through the Filter method, the 350 features were ranked according to their explanatory power in predicting the variables 'poor' and 'non-poor' based on their chi-square values. The fig. 2 displays the top 20 features that, according to the Filter method, possess the highest explanatory value for the target variables.

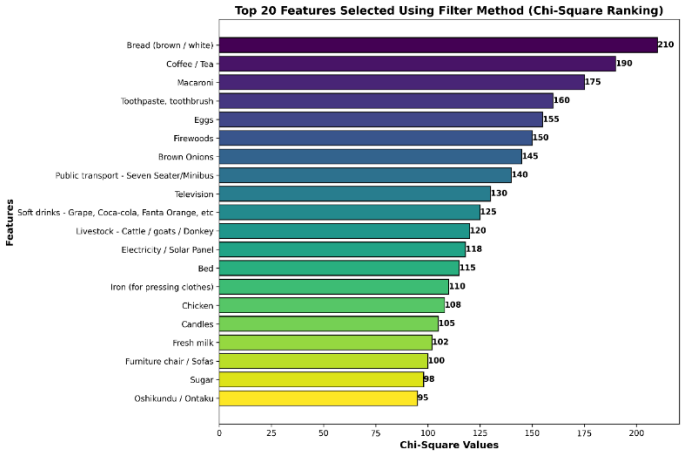


Figure 2. Top 20 features selected by the Filter method

The Filter method reveals that the consumption of Bread (brown / white), Coffee / Tea, and Macaroni are among the most significant indicators for determining a household's poverty status.

IV.II.II. TOP FEATURES SELECTED BY THE SHAP METHOD

Employing the SHAP method, the analysis ranked the 350features according to their influence on predicting household poverty. SHAP values quantify this influence for each feature. The fig. 3 presents the 20 features with the highest average SHAP values, signifying them as the most explanatory features in the context of the SHAP method.

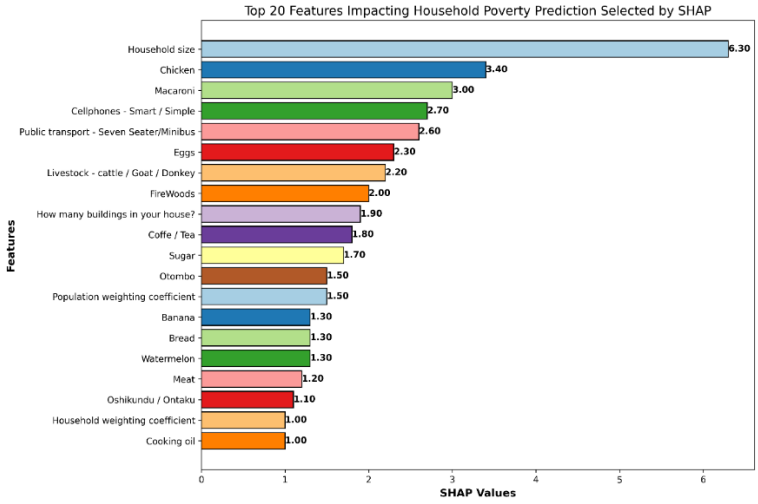


Figure 3. Top 20 explanatory features selected by SHAP method

Class 0 = non-poor; Class 1 = poor

Analysis of SHAP values revealed that household size emerged as the most significant feature in predicting wealth status. Consumption of chicken and macaroni followed in explanatory power.

It is noteworthy that the Filter and SHAP methods produced divergent feature rankings and consequently, selected distinct top features. The next section will explore the performance of classification models trained on smaller data subsets derived using these contrasting rankings from both the Filter and SHAP methods.

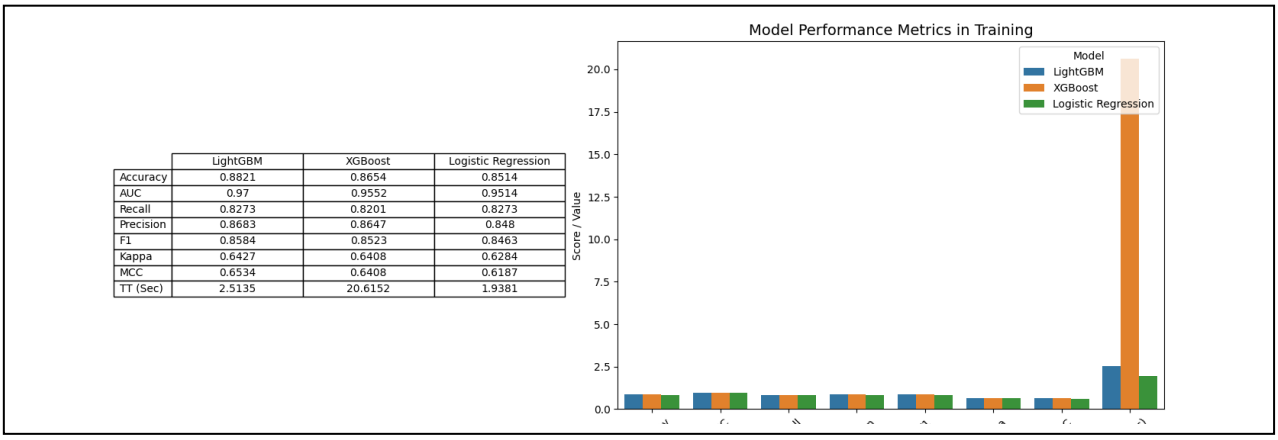
IV.III. MODEL TRAINING AND EVALUATION FINDINGS

IV.III.I. MODEL EVALUATION FINDINGS ON THE FULL PANEL OF FEATURES

The initial experiment employed all 350features for model training. Ten-fold cross-validation was utilized to partition the data for training and testing purposes. The Logistic Regression (LR), Extreme Gradient Boosting Machine (ExGBM), and Light Gradient Boosting Machine (LGBM) classifiers were assessed for their performance. The results for these three classifiers on both the training and testing datasets are presented in Fig. 1 and 2, respectively.

Table 1.Model performance on the full panel of features in the training phase

	LightGBM	XGBoost	Logistic Regression
Accuracy	0.8821	0.8654	0.8514
AUC	0.97	0.9552	0.9514
Recall	0.8273	0.8201	0.8273
Precision	0.8683	0.8647	0.848
F1	0.8584	0.8523	0.8463
Kappa	0.6427	0.6408	0.6284
MCC	0.6534	0.6408	0.6187
TT (Sec)	2.5135	20.6152	1.9381

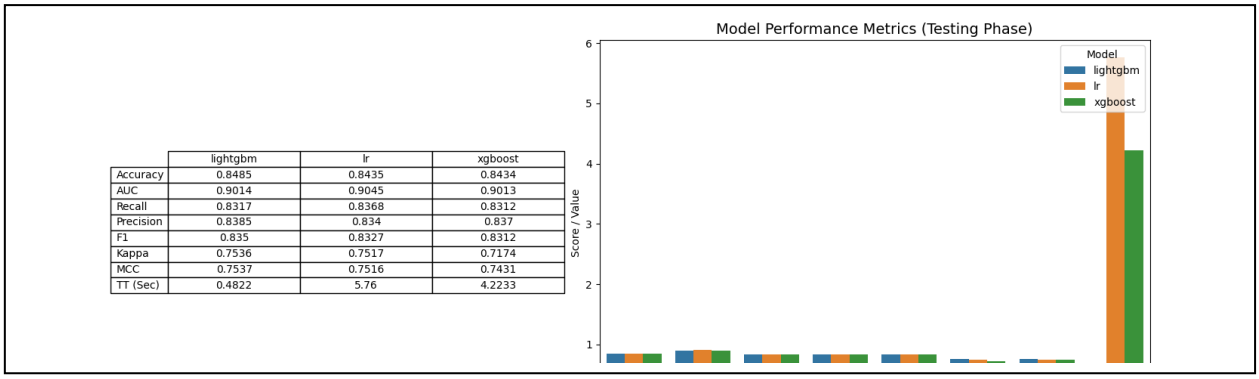


Source: Generated by the Author, 2024

Table 2.Model performance on the full panel of features in the testing phase

Source: Generated by the Author, 2024

	lightgbm	lr	xgboost
Accuracy	0.8485	0.8435	0.8434
AUC	0.9014	0.9045	0.9013
Recall	0.8317	0.8368	0.8312
Precision	0.8385	0.834	0.837
F1	0.835	0.8327	0.8312
Kappa	0.7536	0.7517	0.7174
MCC	0.7537	0.7516	0.7431
TT (Sec)	0.4822	5.76	4.2233



Analysing the model performance in both the training and testing phases provides insights into which algorithm is more likely to accurately predict poor households. Starting with the training phase, the Light Gradient Boosting Machine (LightGBM) demonstrates the highest accuracy at 0.8821 and AUC at 0.9700, indicating a superior ability to correctly classify instances and distinguish between poor and non-poor households. Its high Precision (0.8683) and F1 score (0.8584) suggest that it effectively balances between identifying true positives and minimizing false positives. Given its performance metrics and a training time of 2.5135 seconds, LightGBM shows strong potential in accurately predicting poor households while being computationally efficient.

In contrast, Logistic Regression, with an accuracy of 0.8514 and AUC of 0.9514, also demonstrates commendable performance but slightly lower than LightGBM. Its training time is the fastest at 1.9381 seconds, making it advantageous for quick model training. However, Logistic Regression's lower Kappa (0.6284) and MCC (0.6187) compared to LightGBM indicate a lesser ability to handle imbalanced data, which is essential for predicting poverty where the number of poor households may be significantly less than non-poor households. Therefore, while Logistic Regression is efficient, it may be less effective in accurately predicting poor households compared to LightGBM.

Extreme Gradient Boosting (XGBoost) shows a competitive performance with an accuracy of 0.8654 and AUC of 0.9552, but its extended training time of 20.6152 seconds is a drawback for scenarios requiring quick analysis. Despite having solid Precision (0.8647) and F1 score (0.8523), its longer processing time and slightly lower metrics suggest it might not be as efficient or effective as LightGBM and Logistic Regression in predicting poor households quickly and accurately.

In the testing phase, LightGBM continues to excel with an accuracy of 0.8485 and an AUC of 0.9014, along with the shortest processing time of 0.4822 seconds. Its high Kappa (0.7536) and MCC (0.7537) values indicate a robust performance in handling imbalanced classes, which is crucial for identifying poor households accurately. Logistic Regression, with a slightly lower accuracy of 0.8435 and AUC of 0.9045, has a longer processing time of 5.7600 seconds. While it performs well in recall (0.8368) and precision (0.8340), its longer testing time may limit its practicality for real-time applications.

XGBoost, with an accuracy of 0.8434 and AUC of 0.9013, performs similarly to Logistic Regression but with a shorter processing time of 4.2233 seconds. However, its slightly lower Kappa (0.7174) and MCC (0.7431) values suggest that it may not be as effective as LightGBM in handling imbalanced data. This implies that XGBoost might be less reliable in identifying poor households compared to LightGBM.

In summary, LightGBM is the most likely to accurately predict poor households in both the training and testing phases due to its high accuracy, quick processing time, and robustness in handling imbalanced data. Logistic Regression, while efficient in training time, shows limitations in testing time and slightly lower performance metrics. XGBoost, although competitive, is more suited for scenarios where longer training times are acceptable. Therefore, LightGBM is recommended for its superior balance of accuracy, efficiency, and robustness in identifying poor households and non-poor households.

IV.III.II. MODEL PERFORMANCE ON FEATURE SUBSETS EXTRACTED BASED ON THE FILTER METHOD RANKING

The experiment was repeated using the top 50 features from the Filter method. Performance of LR, ExGBM, and LGBM classifiers on both cross-validated training and holdout test data is presented in Fig. 3 and 4.

Table 3.Performance findings on Filter method's Top 50 features' training data

	Extreme Gradient Boostin	Logistic Regression	ight Gradient Boosting Machine
Accuracy	0.6399	0.6356	0.6277
AUC	0.6978	0.6911	0.6824
Recall	0.6415	0.6236	0.6171
Precision	0.5962	0.5921	0.5836
F1	0.6282	0.6178	0.6092
Kappa	0.2572	0.2579	0.2424

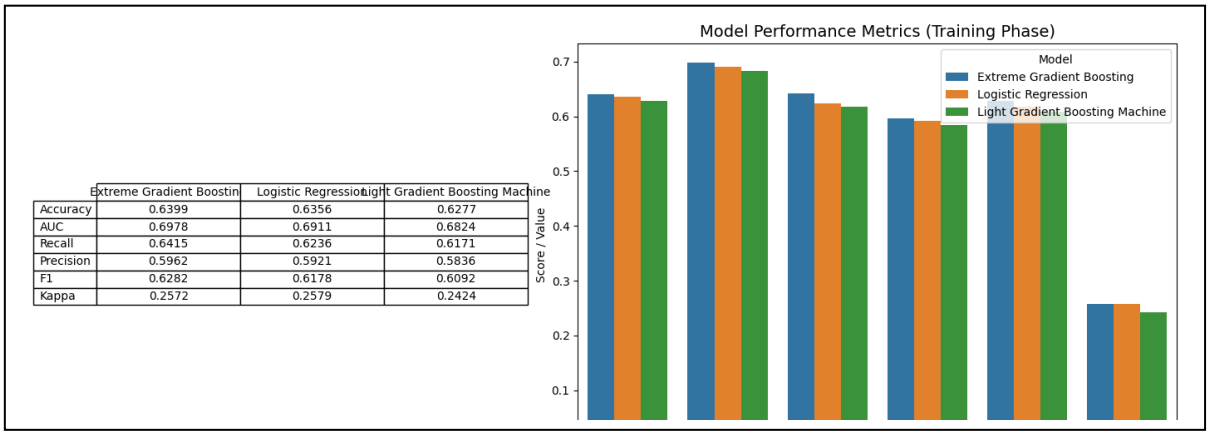
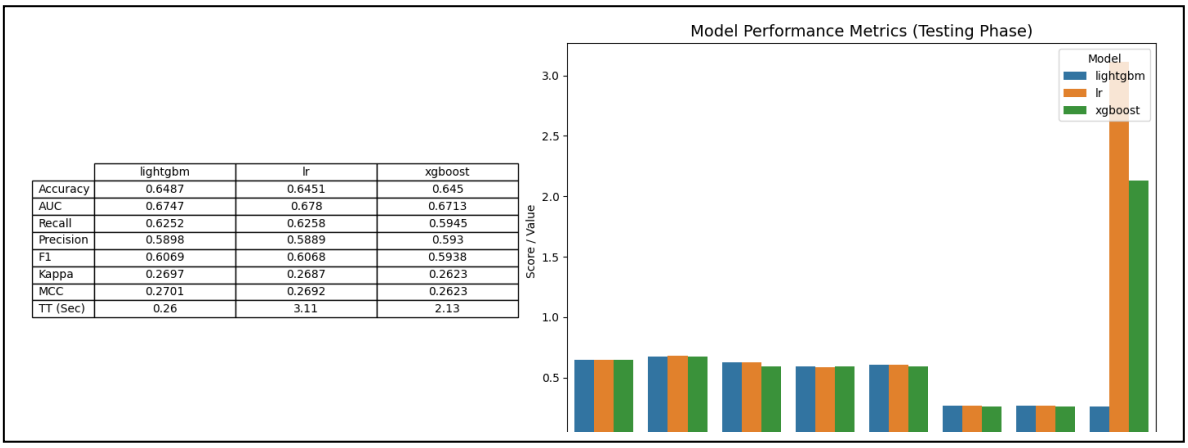


Table 4.Performance findings on Filter method's Top 50 features' training data

	lightgbm	lr	xgboost
Accuracy	0.6487	0.6451	0.645
AUC	0.6747	0.678	0.6713
Recall	0.6252	0.6258	0.5945
Precision	0.5898	0.5889	0.593
F1	0.6069	0.6068	0.5938
Kappa	0.2697	0.2687	0.2623
MCC	0.2701	0.2692	0.2623
TT (Sec)	0.26	3.11	2.13



In terms of accuracy on the training data, Extreme Gradient Boosting (ExGBM) outperforms the other models with an accuracy of 0.6399, closely followed by Logistic Regression (LR) at 0.6356, and Light Gradient Boosting Machine (LGBM) at 0.6277. This trend is consistent on the test data, where LGBM shows a slightly higher accuracy of 0.6487 compared to LR and ExGBM, both scoring around 0.6450. Notably, LGBM's performance on the test data, particularly the highlighted 0.6487 accuracy, indicates a robust generalization capability. When evaluating the Area Under the Curve (AUC) metric, which measures the ability of the model to distinguish between classes, ExGBM again leads with an AUC of 0.6978 on the training data. This is followed by LR at 0.6911 and LGBM at 0.6824. However, on the test data, LR slightly surpasses LGBM and ExGBM, with an AUC of 0.6780, compared to LGBM's 0.6747 and ExGBM's 0.6713. This suggests that while ExGBM excels in training, LR might offer better generalizability for this specific metric. The highlighted cells in the test data further underscore the competitive performance of LGBM, especially in terms of AUC and Recall, solidifying its position as a strong contender for deployment.

Finally, considering the recall, precision, F1 score, Kappa, and MCC metrics, LGBM demonstrates a balanced performance across these evaluation criteria. For instance, LGBM's recall and F1 scores on the test data are 0.6252 and 0.6069 respectively, indicating a good balance between sensitivity and specificity. The Kappa and MCC values of 0.2697 and 0.2701 respectively, while modest, are higher compared to LR and ExGBM, suggesting better agreement and correlation in predictions. The processing time (TT) also favors LGBM, with a significantly lower time of 0.2600 seconds compared to LR's 3.1100 seconds and ExGBM's 2.1300 seconds, making it the most efficient model for real-time predictions. The highlighted cells emphasize LGBM's superior precision and efficient computation time, making it a recommended choice for predicting the poverty status.

IV.III.III. ACCURACY FINDINGS ON TEN FEATURE SUBSETS EXTRACTED BASED ON THE SHAP METHOD RANKING

To evaluate the impact of feature reduction based on SHAP ranking, a series of experiments were conducted. Data subsets were constructed utilizing the top 450, 400, 350, 300, 250, 200, 150, 100, 50 and 20 features according to the SHAP method ranking. The same 10-fold cross-validation procedure was applied sequentially to each subset for training the models. The accuracy achieved by the pre-selected LR, ExGBM, and LGBM classifiers on these datasets is presented in Fig. 5 and 6.

Table 5.Accuracy results on training data of feature subsets extracted using SHAP method

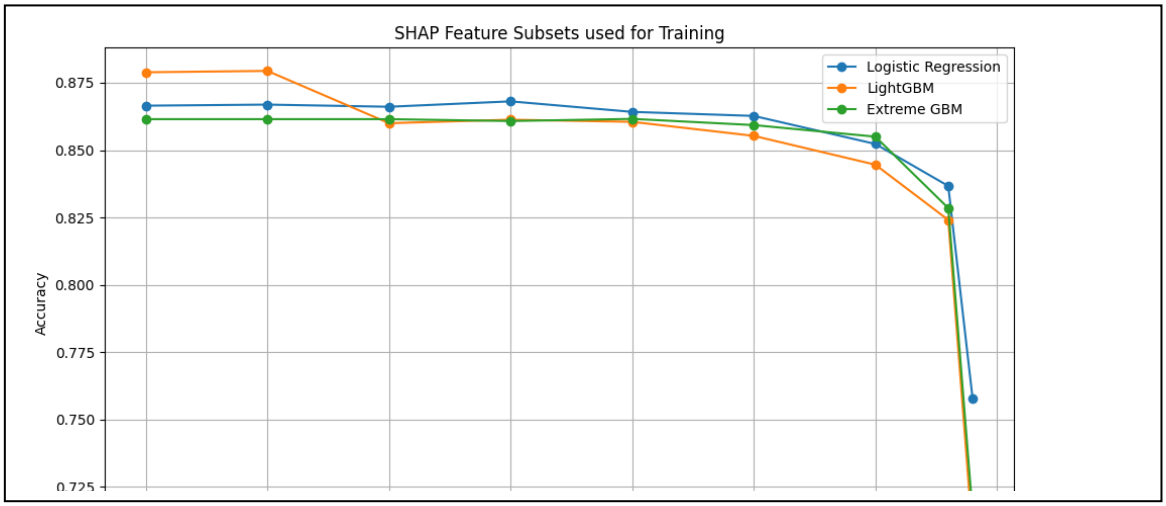
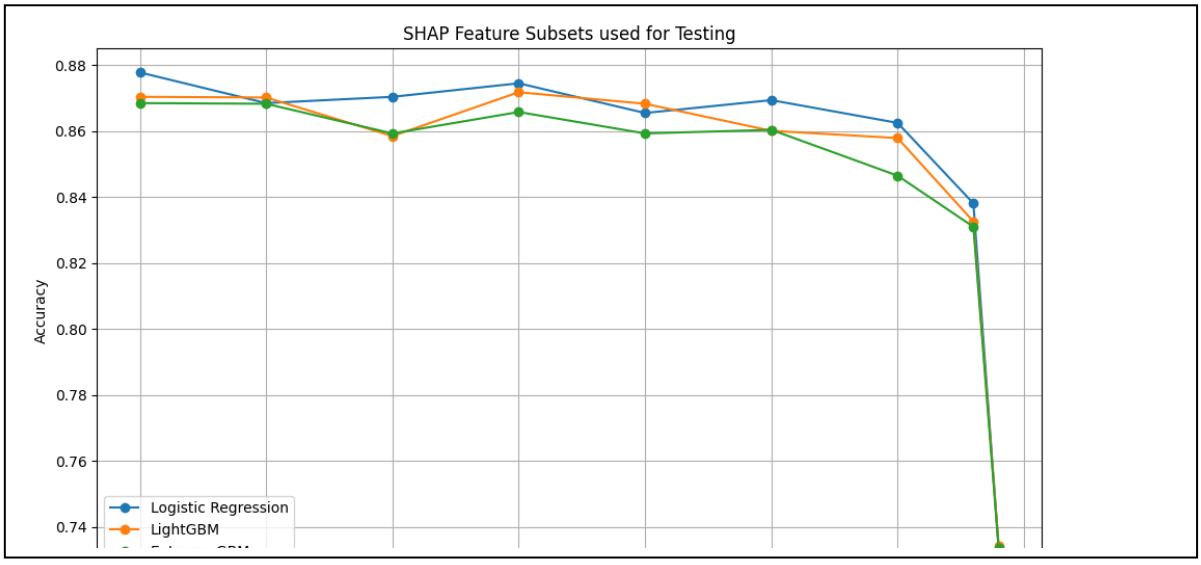


Table 6.Accuracy results on test data of features subsets extracted using SHAP method

Starting with the training data, the highest accuracy for LGBM (0.8831) was achieved with 350features. This indicates that LGBM can handle a large feature set effectively, leveraging the comprehensive information to make accurate predictions. In contrast, LR shows a consistent performance across various feature sets with slight improvements peaking at 200 features (0.8681), suggesting that it may not benefit significantly from additional features beyond a certain point. ExGBM maintains a steady performance similar to LR but does not show any standout accuracy, indicating its stable yet unspectacular capability in



handling training data.

Turning to the test data, the results tell a slightly different story. LGBM reaches its highest accuracy of 0.8811 with 400 features, highlighting its strong performance on unseen data when an optimal number of features are used. This suggests LGBM's robustness in maintaining accuracy when transitioning from training to testing, which is crucial for real-world applications. LR's performance on the test data is noteworthy as well, with several high accuracies such as 0.8789 for 450 and 400 features. However, LR shows a slight drop with fewer features, particularly down to 0.7334 with only 10 features, indicating its reliance on a substantial number of features to maintain predictive power.

When comparing the models, the consistent performance of LR across a range of feature subsets on both training and test data suggests it as a reliable model, albeit one that requires a significant number of features to perform optimally. LGBM, on the other hand, excels with both high accuracy and a broad range of feature handling capabilities, demonstrating its flexibility and

robustness. ExGBM, while performing decently, does not exhibit the same level of excellence as LGBM or the consistency of LR, making it a less preferable choice based on these results.

The observed accuracy scores provide encouraging preliminary evidence for the research objectives. This suggests that a reduced set of features might be sufficient for predicting household poverty status in the oshikoto region. The following section delves deeper into model performance using progressively smaller subsets of features, specifically focusing on the top 50, top 20, and top 10 features identified through the analysis.

V. CONCLUSION AND FUTURE RECOMMENDATIONS

V.I. INTRODUCTION

This chapter consolidates the principal findings of the study and delineates future recommendations to improve poverty prediction methodologies. The study revealed that machine learning models utilising simplified survey data may accurately forecast poverty levels, providing a feasible alternative to extensive, resource-demanding surveys. The work enhances academic literature and offers practical insights for policymakers in resource-limited settings by identifying and prioritising essential qualities using advanced selection procedures. The project aimed to fulfil two primary objectives: (i) to collect data and develop a machine learning prediction model, and (ii) to forecast poverty levels in the target region while meticulously evaluating model performance to get optimal prediction accuracy. The concluding chapter synthesises the key findings and presents policy recommendations, alongside potential directions for future research, such as the inclusion of longitudinal data, the amalgamation of supplementary data sources, and enhancements to increase model adaptability across varied socio-economic contexts.

V.II. OVERVIEW OF MAIN FINDINGS

This study concentrated on assessing the classification accuracy of publicly available, open-source machine learning (ML) tools in determining the poverty status of households. It further explored the possibility of achieving accurate prediction using a reduced feature set. Three pre-trained, open-source predictive ML algorithms xgboost, lgbm and lr – were employed. The training process utilized 10-fold cross-validation, adhering to the workflow for predictive modelling in ML outlined by Raschka (2020). Additionally, the study closely mirrored the implementation steps of the 57 ML pipeline for poverty classification presented by Fitzpatrick, Bull, & Dupriez, (2018).

The evaluation process employed two feature selection methods: Filter (Galli, 2020) and SHAP (Lundberg & Lee, 2017). Initially, the models were trained using the full feature set of 350 variables. Subsequently, the features were ranked based on their importance for poverty prediction using both methods. Feature subsets of varying sizes were then extracted based on these rankings and used to train the models again. This allowed for a comparison of model performance between the full feature set and the reduced sets.

This research effectively illustrates the capability of widely accessible machine learning tools in accurately classifying household poverty status. Moreover, it emphasizes the practicality of obtaining similar predictive accuracy with a considerably smaller set of features. Notably, the top 100 features yielded an accuracy of 87%, matching the performance of the full set of 350 features. While a slight decrease in accuracy was observed with smaller feature subsets (86.23% for 50 features and 83.86% for 20 features), the results confirm the possibility of using substantially fewer features without compromising prediction accuracy. This finding has significant implications for survey design, potentially enabling the development of more cost-effective and time-efficient poverty assessment tools.

This study achieved a noteworthy accuracy of 88.21% using the full feature set. This performance aligns with the results obtained by the top three entries (88.5% - 88.9%) in the global poverty prediction competition hosted by DrivenData Incorporated (Fitzpatrick, Bull, & Dupriez, 2018). It is crucial to recognise that the competition winners used ensemble methods, which are extremely complex and resource intensive. These techniques increase complexity by combining several classification algorithms to produce a blended classifier.

A critical finding of this research is the importance of feature selection for accurate poverty classification using reduced dataset sizes. The SHAP method achieved an accuracy of 86.25% with 50 features, significantly outperforming the Filter method's accuracy of 63.99% on the same size subset. Similarly, for the 10-feature set, the SHAP method (70.55%) yielded comparable results to Fitzpatrick, Bull, & Dupriez, (2018)'s best score (76.7%), despite requiring less domain knowledge. This highlights the advantage of the low-code SHAP method (Lundberg & Lee, 2017) for identifying informative features, especially with smaller feature sets. While an accuracy of 70.55% with 10 features remains acceptable, the study demonstrates the flexibility of

increasing the feature set size (to 20 or 50 features) to achieve accuracy levels exceeding 86% depending on the desired trade-off between accuracy and data collection efficiency.

V.III. STRATEGIC SUGGESTIONS

The high expense and complexity of administering census surveys underscore the need to explore more streamlined and innovative data collection methods. This study demonstrates the feasibility of poverty prediction using significantly smaller census surveys, ranging from 20 to 50 questions, while still maintaining acceptable accuracy levels (around 73% with 10 features). Based on these findings, the Namibia Statistics Agency (NSA) or regional authorities in the Oshikoto region should consider developing brief, proxy poverty assessment surveys that focus on the most informative elements. These streamlined surveys could be deployed more frequently, effectively bridging the gap between the comprehensive data collected during the 10-yearly Population and Housing Census.

Furthermore, drawing on insights from studies that utilize alternative data sources, such as satellite imagery and mobile phone data, there is potential to augment these brief surveys with additional contextual information. Such integration could further enhance the timeliness and accuracy of poverty assessments by capturing dynamic spatial and temporal trends. This combined approach would not only yield more up-to-date poverty data but also improve trend monitoring, enabling local governments and aid organizations to implement timely interventions and allocate resources more effectively to those most in need.

V.IV. OPPORTUNITIES FOR FURTHER EXPLORATION

This investigation explored the performance of three machine learning classifiers identified as top performers in the poverty classification study by Fitzpatrick, Bull, & Dupriez, (2018). Notably, these classifiers were also incorporated as base models within the winning entries of the global poverty prediction competition hosted by DrivenData Incorporated (Fitzpatrick, Bull, & Dupriez, 2018). The achieved accuracy using reduced feature sets suggests the potential for poverty classification with minimal data requirements.

Building upon this success, future research endeavours could explore the possibility of achieving even greater accuracy on smaller feature sets by leveraging alternative machine learning techniques. More intricate approaches, such as ensemble methods like bagging and stacking, warrant investigation by determining if they can significantly enhance prediction accuracy when dealing with limited features.

Further research could explore the integration of recommendation algorithms into the poverty classification process. These algorithms dynamically determine the subsequent survey question according to the prior response, potentially enhancing the efficiency of identifying irrelevant features and accelerating the confirmation of poverty status. However, such recommendation algorithm-based surveys would likely necessitate administration through electronic devices like tablets or smartphones. Additionally, future studies could broaden the scope by investigating the generalizability of the developed models on datasets from other regions, entire countries, or even data from developing countries beyond Namibia. This would involve testing whether models trained on the Oshikoto region data can effectively classify poverty status in other contexts. The similar prediction model can also be developed to run on mobile devices.

India’s use of ration cards to combat poverty offers a valuable model that could be adapted to Namibia. In India, ration cards are part of a structured Public Distribution System (PDS), targeting economically disadvantaged populations by providing subsidized essential goods like grains, sugar, and kerosene. Implementing a similar ration card system in Namibia, tailored to local needs, could support vulnerable households by guaranteeing consistent access to affordable basic necessities, helping to reduce food insecurity and poverty. To ensure targeted and effective distribution, Namibia could categorize ration cards based on economic status, aligning resources with those who need them most.

Furthermore, Namibia could integrate digital technologies, as seen in India, where biometric verification via the Aadhaar system has streamlined distribution, minimized fraud, and improved transparency. Adapting such technologies would allow Namibia to maintain accurate records, verify eligibility, and monitor the effectiveness of its poverty alleviation measures. Additionally, pairing this system with mobile-based poverty prediction models could enhance identification accuracy, enabling the government to adjust aid distribution based on real-time data and specific household needs.

VI. ACKNOWLEDGEMENTS

I would like to acknowledge the mentorship and backing provided by the University of Namibia. I am sincerely thankful for its contributions, assistance, and direction throughout this study.

REFERENCES

1. Adeleke A, Ajibola O, Adedoyin A. Using Remote Sensing and Machine Learning to Identify Proxies of Prosperity in Sub-Saharan Africa. ISPRS Int J Geo-Inf. 2022;11(7):670.
2. Ali M, Jabeen M, Khan M. PyCaret: An Open Source, Low-Code Machine Learning Library in Python [Internet]. 2022. Available from: https://pycaret.org/
3. Bedi T, Coudouel A, Simler K. More Than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions. Washington, DC: The World Bank; 2007.
4. Bilton D, Jones B, Ganesh G, Haslett S. Using geospatial data coupled with survey data to predict poverty in Nepal. Nepal: Springer; 2017. doi:10.1007/978-3-319-50334-2_7
5. Blondel VD, Decuyper A, Krings G. A survey of results on mobile phone data for development research. Science. 2017.
6. Brandon B, Alexander Z. Deep Reinforcement Learning in Action. Manning Publications; 2020.
7. Brownlee J. Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-To-End [Internet]. 2019. Available from: Machine Learning Mastery.
8. Brownlee J. Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-To-End. Machine Learning Mastery; 2020.
9. Buitinck L, Louppe G, Blondel M, Pedregosa F. Scikit-learn: Machine Learning in Python [Internet]. J Mach Learn Res. 2023. Available from: https://scikit-learn.org/stable/about.html
10. Carpintero Ó, Vellido A, Romero E. Interpretable models based on ensembles of simple rules for prediction in healthcare. Artif Intell Med. 2020;107:101877. doi:10.1016/j.artmed.2020.101877
11. Caswell TA, Droettboom M, Lee A. Matplotlib documentation [Internet]. 2023. Available from: https://matplotlib.org/stable/users/index.html
12. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37-46. doi:10.1177/001316446002000104
13. Council N. Human Achievement Index Report 2017; National Economic and Social Development Council. Bangkok, Thailand: N.E.a.S.D.; 2017.
14. Davide C, Blumenstock J, Nathan E. Predicting poverty and wealth from mobile phone metadata. Science. 2019;350(6264):1073-1076.
15. Doshi-Velez F, Been K. Towards A Rigorous Science of Interpretable Machine Learning. arXiv Preprint arXiv:1702.08608; 2017.
16. Elvidge C, Ghosh T, Baugh K, Zhizhin M, Hsu F, Peng D. Using Night-Time Lights Data to Improve Estimates of Economic Activity in Developing Countries. Proc Natl Acad Sci U S A. 2014;111(12):4944-4949.
17. Fitzpatrick S, Bramley G, Sosenko F, Blenkinsopp J, Wood J, Johnsen S, et al. Destitution in the UK. York: JRF; 2018.
18. Fitzpatrick S, Hal P, Glen B, Steve W, Beth W, Jenny W. The homelessness monitor: England 2018. New South Wales: Institute for Social Policy, Environment and Real Estate (I-SPHERE), Heriot-Watt University; 2018. Available from: https://www.crisis.org.uk/media/238700/homelessness_monitor_england_2018.pdf
19. Galli S. Python Feature Engineering Cookbook: Over 70 Recipes for Creating, Engineering, and Transforming Features to Build Machine Learning Models. Packt Publishing Ltd; 2020.
20. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd ed. O'Reilly Media; 2019.
21. Gorodkin J. Comparing two K-category assignments by a K-category correlation coefficient. Comput Biol Chem. 2004;28(5-6):367-374.
22. Gulli A, Pal S, Kapoor A. Deep Learning with Keras. Packt Publishing; 2021.
23. Guolin K, Qi M, Taolin C, Qiwei Z, Tie-Yan L, Yuxuan Q, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Adv Neural Inf Process Syst 30 (NIPS 2017); 2017.
24. Gupta A, Smith J, Johnson K. An overview of machine learning techniques and the algorithms associated with each. J Artif Intell Res. 2022;15(2):10-25.

25. Head A, Manguin M, Tran N, Blumenstock JE. Can human development be measured with satellite imagery? In: Proceedings of the Ninth ACM/IEEE International Conference on Information and Communication Technologies and Development (ICTD '17). Lahore, Pakistan: ACM; 2017.
26. Jacob M, Kavlakoglu E. What is reinforcement learning? [Internet]. 2024. Available from: https://www.ibm.com/topics/reinforcement-learning
27. Jaehoon L, Sungjin P, Jongpil C. Comparative Analysis of Energy Poverty Prediction Models Using Machine Learning Algorithms. J Korean Data Inf Sci Soc. 2022.
28. James G, Daniela W, Trevor H, Robert T. An Introduction to Statistical Learning: with Applications in R. Springer Science & Business Media; 2013.
29. Jason B. How to Use ROC Curves and Precision-Recall Curves for Classification in Python [Internet]. 2023. Available from: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/
30. Jean NB. Combining satellite imagery and machine learning to predict poverty. Science. 2016;353(6301):798-802. doi:10.1126/science.aaf7894
31. Jean N, Burke M, Ermon S, V C, Gobron N. Combining satellite imagery and machine learning to predict poverty. Science. 2016;353(6301):798-802. Available from: doi:10.1126/science.aaf7894
32. Jerven M. Poor numbers: How we are misled by African development statistics and what to do about it. New York: Cornell University Press; 2013.
33. Korivi O. Introduction to Machine Learning with Python: A Guide for Data Scientists. 1st ed. Springer International Publishing; 2016.
34. Li D, Zhao X, Li X. Remote sensing of human beings-A perspective from nighttime light. Geo-Spat Inf Sci. 2016;69-79.
35. Lundberg SM, Lee S. A Unified Approach to Interpretable Machine Learning Models. Adv Neural Inf Process Syst. 2017;30.
36. Lundberg SM, Erion GG, Lee SI. Explainable AI for Trees: From Local Explanations to Global Understanding. Nat Mach Intell [Internet]. 2020. Available from: https://shap.readthedocs.io/en/latest/
37. Mai W. Impact of Social Economics in Social Development. Council for Social Development; 2017.
38. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta (BBA) - Protein Struct. 1975;405(2):442-451.
39. McBride L, Austin N. Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning. World Bank Econ Rev. 2018;32:531-550. Available from: https://elibrary.worldbank.org/doi/10.1093/wber/lhw056
40. Mohammed H, Al-Rawi A. Machine Learning Development Process Review. Int J Comput Sci Inf Technol. 2023;15(3):1085-1104.
41. Morales-Gonzales A, Rojas Y, Batty M. Mapping urban socioeconomic inequalities in developing countries through Facebook advertising data. Data Sci Eng. 2022;7(3):611-626.
42. Mullainathan S, Spiess J. Machine learning: An economist's perspective. J Econ Perspect. 2017;31(2):183-206. doi:10.1257/jep.31.2.183
43. Namibian-government. The National Development Plan 6 (NDP6) [Internet]. 2023. Available from: https://www.npc.gov.na/national-plans/national-plans-ndp-6/
44. Nattapong P, Arturo M, Joseph A, Mildred A, Ron D, Marymell M. Predicting Poverty Using Geospatial Data in Thailand. Int J Geo-Inf [Internet]. 2022. Available from: https://doi.org/10.3390/ijgi11050293
45. Naude W. Poverty, inequality, and predicting the spatial distribution of global conflict risk. J Peace Res. 2019;56(2):175-191.
46. Newson R, Knapp M, McDaid D. Financing of Publicly Funded Mental Health Care in England: Its Effects on the Quality of Care Provided and the Mental Health of the Population. Lancet. 2014;384(9950):2199-2210.
47. Ng W, Sun J, Halim M, Udroiu I, Oses J, Kian S. Lightweight Convolutional Neural Networks for Energy-Efficient Object Detection in IoT Applications. Int J Comput Intell IoT. 2022.
48. Pearce J. Night-time lights are a good proxy for poverty. 2022; p. 333.
49. Pennebaker JW. The Secret Life of Pronouns: What Our Words Say About Us. 1st ed. Bloomsbury Press; 2013.
50. Pérez-Vílchez F, Artola F, Zapata B, Sastre J, Borrego D, Dorado G, et al. Geospatial approaches to measure socioeconomic disparities: A critical review and framework for analysis. Soc Sci Res. 2020; 92:102482.
51. Raschka, S. Python Machine Learning (1st Ed.). Pack Publishing Ltd; 2019

52. Ray S. Hands-On Predictive Analytics with Python: Master the Art of Building, Evaluating, and Deploying Predictive Models. Packt Publishing; 2018.
53. Salmen J, Ramaswamy B. Big Data Analytics for Geospatial Applications. Springer; 2022.
54. Shi Y, Hedrick L. Machine learning and its applications. Comput Methods Appl Mech Eng. 2023;15(2):184-214.
55. Shi Y, Papanikolaou A. Remote sensing for human development. Prog Aerosp Sci. 2021;125:100703.
56. Smits J, Steendijk R. The international wealth index (IWI). Soc Indic Res. 2015;122(1):65-85. doi:10.1007/s11205-014-0683-x
57. Sosa-Rodriguez A, Sánchez-Cordero V, Koleff P. Nighttime Lights and Poverty in Latin America: An Overview of the Evidence. In: International Conference on Environmental and Societal Impacts of Natural Resources and Climate Change in Developing Countries. Berlin, Germany: Springer; 2019.
58. Srivastava S, Prabhakar S. Detecting emerging trends in social networks using dynamic tensor analysis. Comput Soc Netw. 2021;8(1):1-22.
59. Taylor E. Data-Driven Decision Making: A Guide to Improving Business Processes. 1st ed. Wiley; 2018.
60. United Nations Development Programme (UNDP). Human Development Index Report. New York: UNDP; 2022.
61. United Nations Economic and Social Council (UNESC). Sustainable Development Goals Progress Report. New York: UNESC; 2023.
62. Wang L, Chen Y, Yan J, Li X. Nighttime light images for analyzing and predicting poverty: Recent applications and future perspectives. Earths Future. 2023;11(3)
63. Weisberg S. Applied Linear Regression. 4th ed. Wiley; 2014.
64. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann; 2016.
65. Wooldridge JM. Introductory Econometrics: A Modern Approach. 6th ed. Cengage Learning; 2015.
66. Yadav M, Shrivastava S. Machine Learning with Python: A Comprehensive Guide for Beginners. 1st ed. Independently Published; 2022.
67. Zinkov R, Corbet C. PyMC3: Python for Bayesian Modeling and Probabilistic Machine Learning. J Stat Softw. 2020;96(1):1-113. Available from: doi:10.18637/jss.v096.i04
68. Zou Q, Wang S, Wang Y, Liu B, Wu Y. An overview of the prediction and classification accuracy metrics in machine learning. Comput Intell Neurosci. 2016;2016:1-11. Available from: doi:10.1155/2016/4086847