Machine Learning-Based System to Predict Poverty - Stricken Households in Rural Areas: The Case of Zimbabwe

Cloudy Mbuwa

School of Information Science and Technology (SIST), Department of Software Engineering (SE), Harare Institute of Technology (HIT)
Harare, Zimbabwe
Email: *cmbuwa80[at]gmail.com*

Abstract: ***In most countries in the world, poverty eradication is the most common subject, but what differs is the methods and techniques applied in identifying those who are in poverty. Different techniques which include night satellite images, surveys and censuses are among the methods used to try and identify population who are in different categories of poverty such as Extremely poor, Moderate poor and poor. Levels of poverty cannot be measured using the same indicators. Different indicators and different methodologies are applied across different countries in an effort to identify and classify households in poverty. The poverty measuring indicators vary depending on whether the household is in rural or urban area. The use of obsolete methodologies and data collection and processing technology lead to delays and error prone in the identification and delivery of government assistance to the households who are in extreme poverty. This study aims at developing a machine learning web-based system for classifying poverty-stricken households in rural areas of Zimbabwe. Among the supervised machine learning algorithms that was considered, Logistic Regression algorithm was the best poverty-stricken household classifier. Household Targeting Surveys (HTS) that was done by Ministry of Social Welfare and Zimstat since 2011 up to 2022 was the source of dataset used to train the model. For the purpose of identifying and classifying poverty-stricken rural households in Zimbabwe, indicators such as assets ownership, land ownership, household size, health status of head of household etc. was considered for the classification of households. See Fig. 1 for complete list of features that was extracted and considered. After train Logistic Regression algorithm on the HTS dataset, the model performance was 99% and this was achieved after performing model evaluation techniques such as cross validation.***

Keywords: Feature Engineering, Poverty, Cross-validation, Evaluation Metrics, Regularization

1. Introduction

By 2030, Goal 1 of the Sustainable Development Goals (SDGs) seeks to eradicate all forms of poverty worldwide. Governments, legislators and other organisations have been working to reduce poverty for years. The amount of income and the availability of sufficient basic resources to support sustainable livelihoods are typically determining factors. In fact, it also highlights the lack of access to education, starvation and malnutrition, social injustice, and other necessities of life [1].

According to the Second United Nations Decade for the Eradication of Poverty (2008-2017), which promotes a people-centric strategy in targeting the most vulnerable groups, poverty reduction should be ingrained in national policies and approached from all dimensions, including political, economic, and social ones.

Data from household expenditure and income surveys is used to categorise and quantify the poverty status of households. But conducting such surveys is challenging, time-consuming, and expensive. Even worse, by the time the data are gathered and examined, they are frequently outdated, which implies that policymakers are more likely to base their judgements on outdated information. Machine learning has the potential to fundamentally alter the game by making poverty assessment and identification more affordable and much more in-the-moment [2].

The aims of the study are to produce initial and objective results with regard to building a machine learning model capable of predicting the poverty status of rural households in Zimbabwe. This study is the first of its kind in the country. In this perspective, it is possible to look into the role of machine learning in the fight against poverty in more detail. The identification of intricate relationships between numerous economic indicators and between other dissimilar data sources aids in comprehending the development of society. In this study, data on household characteristics are the subject of the analysis.

On the basis of information regarding household characteristics, a variety of supervised machine learning models are deployed to forecast poverty. Employing supervised machine learning in this area will have a significant impact on policy makers' and decision-making in order to provide financial aid to selected low-income households, including those in poverty, extreme poverty, and moderate poverty.

The goal of this study is to predict the amount of poverty by examination of data on household characteristics using a variety of regression models. Instead of focusing solely on income level, features such as household assets wealth, land ownership, and other living quality circumstances (see appendix) were taken into consideration for this study as the factors of poverty.

Building on existing efforts, data from Rural Housing Target Survey (RHTS) database was used for model building process. The data was first anonymized in order to preserve the secrecy of households. Likewise, historical data collected from Target Household Surveys (HTS 2011-2022)

Paper ID: SR25210190332 DOI: https://dx.doi.org/10.21275/SR25210190332

might contain redundancy, and it can be imbalanced. This may lead to incorrect prediction due to too many data errors such as missing values, imbalanced and redundant data.

In order to minimize the errors, different data handling methods will be applied to ensure the accuracy of the supervised machine learning model. Another aim of this research is to understand the individual features that is important for the classification of rural households according to the level of poverty. Through the use of python libraries such as Pandas, functions such as correlations are used to understand the degree of relationship between predictor variables and the dependent variable. This offers a powerful and insightful measure of the importance of a feature in a model. Thus, we can know the contribution of certain features in a particular proposed poverty prediction model, as a result this can assist in decision-making and policy formulation.

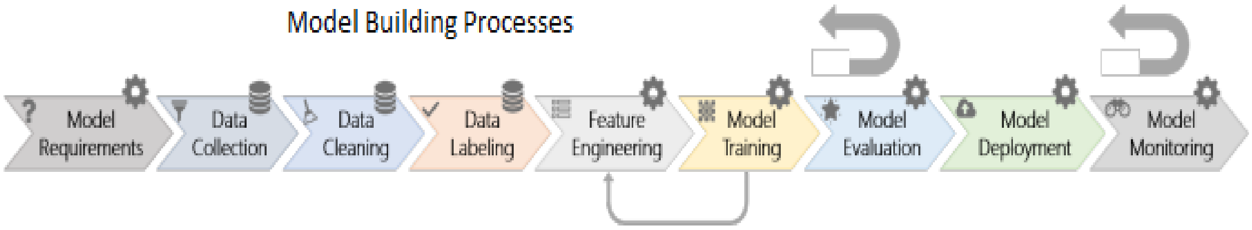
This paper is divided into six chapters: chapter 1 gives an introduction to the study; chapter 2 reviews some related literature, while chapter 3 explains the methodology used in this paper; chapter 4 data presentation and analysis, chapter 5 discuss the results of the study; and chapter 6 summarize, give conclusion, and recommendations of the study.

2. Problem Statement

The stimulus for the study was to identify and implement the best and efficient machine learning algorithm for classifying rural households according to their level of poverty.

3. Material and Method

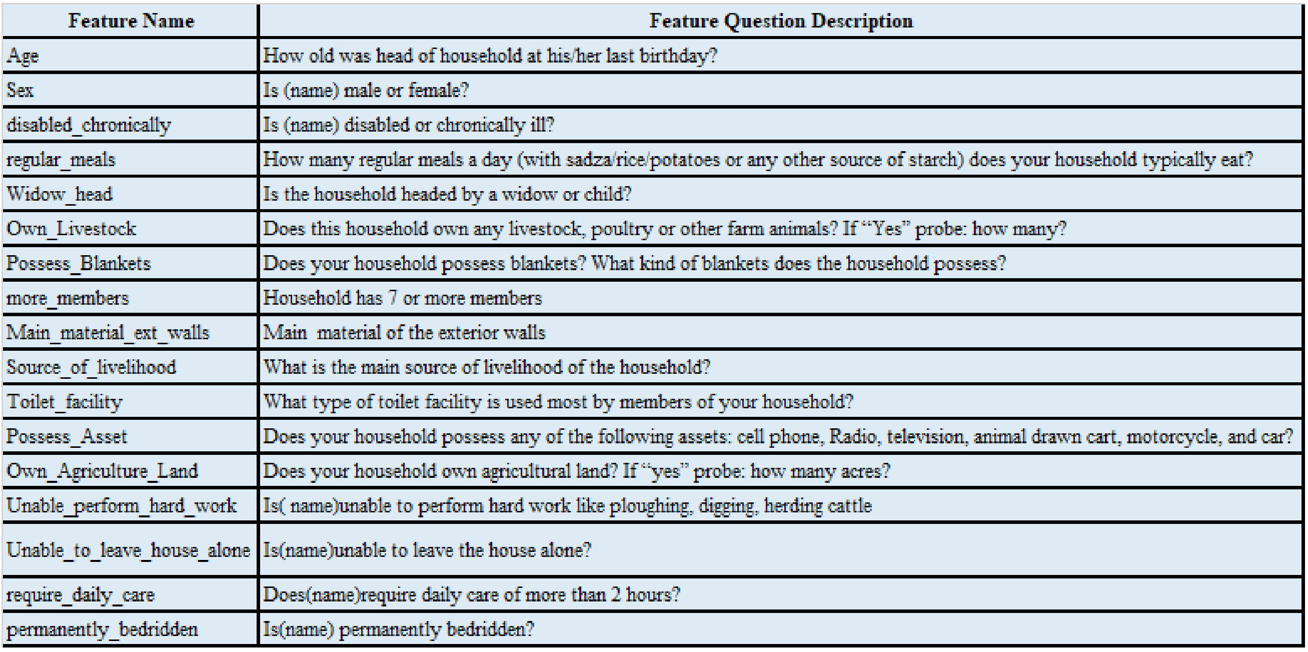
The research methodology of this study is divided into three phases. The dataset phase starts by identifying data examined in this study and analysing its source, details and quantity.



This is followed by the Pre-processing phase which aims to prepare the data for processing. Particularly, this phase includes four tasks, which are data cleaning. feature engineering, Normalisation, and sampling method. Pre-processed data was then used in the third phase to establish a comparative analysis among the three techniques to identify the best machine learning technique.

a) Dataset Description

Table 1: List of Features



Feature Name	Feature Question Description
Age	How old was head of household at his/her last birthday?
Sex	Is (name) male or female?
disabled_chronically	Is (name) disabled or chronically ill?
regular_meals	How many regular meals a day (with sadza/rice/potatoes or any other source of starch) does your household typically eat?
Widow_head	Is the household headed by a widow or child?
Own_Livestock	Does this household own any livestock, poultry or other farm animals? If "Yes" probe: how many?
Possess_Blankets	Does your household possess blankets? What kind of blankets does the household possess?
more_members	Household has 7 or more members
Main_material_ext_walls	Main material of the exterior walls
Source_of_livelihood	What is the main source of livelihood of the household?
Toilet_facility	What type of toilet facility is used most by members of your household?
Possess_Asset	Does your household possess any of the following assets: cell phone, Radio, television, animal drawn cart, motorcycle, and car?
Own_Agriculture_Land	Does your household own agricultural land? If "yes" probe: how many acres?
Unable_perform_hard_work	Is(name)unable to perform hard work like ploughing, digging, herding cattle
Unable_to_leave_house_alone	Is(name)unable to leave the house alone?
require_daily_care	Does(name)require daily care of more than 2 hours?
permanently_bedridden	Is(name) permanently bedridden?

For this study, pre-labelled dataset was used to train the algorithm to identify three different classes of poverty-stricken households in rural areas, and how well it predicts the wealth class of rural households especially for unseen data cases. The dataset used in this study comes from Rural Household Targeting Survey (HTS). The survey was carried out annual since 2011-2018. The main objective of the survey was to identify rural households which require urgent government assistance through the department of Social Welfare. For this study, a total of 33 642 households records were used from different districts.

b) Pre-Processing:

Data transformation was done to the dataset, this includes data cleaning, feature engineering, normalisation, feature selection and sampling methods so that the dataset is suitable for the building of the machine learning model. There are several various data mining tools that can be used for data pre-processing purposes. In this study, the python pandas and Scikit-learn library was used as a tool to perform the pre-processing and machine learning model building task. Scikit-learn library contains various types of machine learning algorithms and operates on an opensource license. Matplotlib library was used to provide various visualization for data analysis

c) Data Cleaning:

Before starting the data cleaning process, data visualization can be utilized to get an overview of the basic pattern of the dataset in a graphical view.

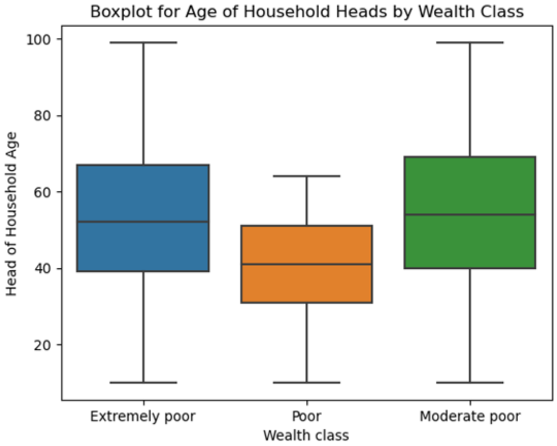


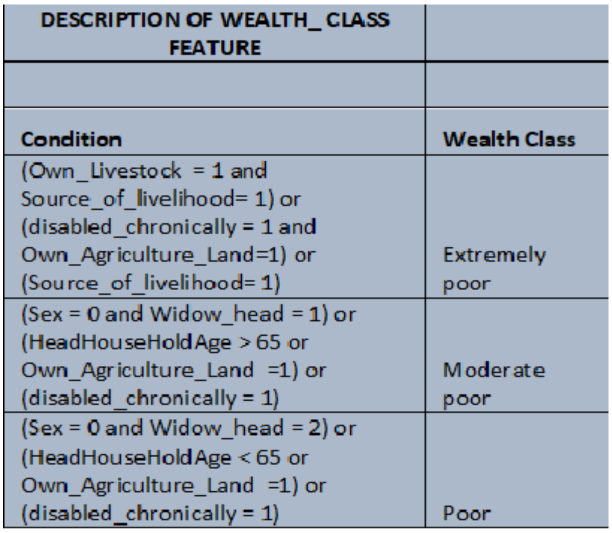
Figure 1

Figure 2 shows boxplot for wealth class and age of head of household head. Boxplot is one of the visualizations that was used to heck for age outliers. Data cleaning activities such as removing outliers, delete duplicates, removing errors or noise and removing invalid data was performed on the data set.

d) Feature Engineering:

The HTS data set does not have a wealth class category. The pre-labelled class for wealth class was generated based on the following threshold as shown in table below. The wealth class category threshold was engineered as described by the table below.

Table 2



DESCRIPTION OF WEALTH_ CLASS FEATURE	
Condition	**Wealth Class**
(Own_Livestock = 1 and Source_of_livelihood= 1) or (disabled_chronically = 1 and Own_Agriculture_Land=1) or (Source_of_livelihood= 1)	Extremely poor
(Sex = 0 and Widow_head = 1) or (HeadHouseHoldAge > 65 or Own_Agriculture_Land =1) or (disabled_chronically = 1)	Moderate poor
(Sex = 0 and Widow_head = 2) or (HeadHouseHoldAge < 65 or Own_Agriculture_Land =1) or (disabled_chronically = 1)	Poor

e) Feature Scaling:

Feature scaling is every crucial step in building machine learning model to ensure that features are within the same scale. Standardization was conducted to transform the data to have a mean of zero and standard deviation of 1. Standardization is also known as Z-score normalization in which properties will have the behaviour of standard normal distribution [3]. The standardization formula:

$$z = \frac{x - \bar{x}}{\sigma}$$

f) Feature Selection:

Feature selection is a process to improve classification accuracy by removing irrelevant and redundant features from the original dataset. The technique is used to reduce the dimensionality of the dataset hence improve model interpretability, learning accuracy and reduce running time [3]. There are three broader techniques for feature selection that is Filter, Wrapper and Embedded. In this study, the filter technique was used i.e feature importance-based technique was applied. Feature importance is an inbuilt class that comes with tree-based classifiers such as decision tree and random forest.

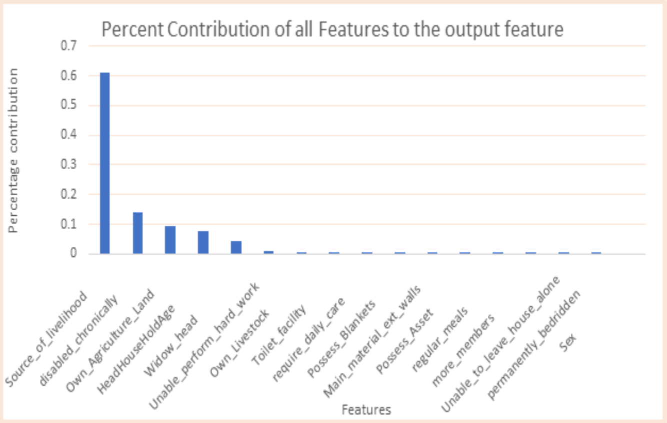


Figure 2

g) Sampling Method Selection:

In the HTS experimental dataset, the number of poor and moderate poor class is dominated by Extremely poor majority class as a result an imbalance dataset. This causes the classifiers to get biased towards the majority class. There are techniques to deal with imbalance dataset such SMOTE, etc. In this study, technique of dealing with imbalance dataset was used. The solution to imbalance dataset was to use weighted average F-1 score as a performance metric.

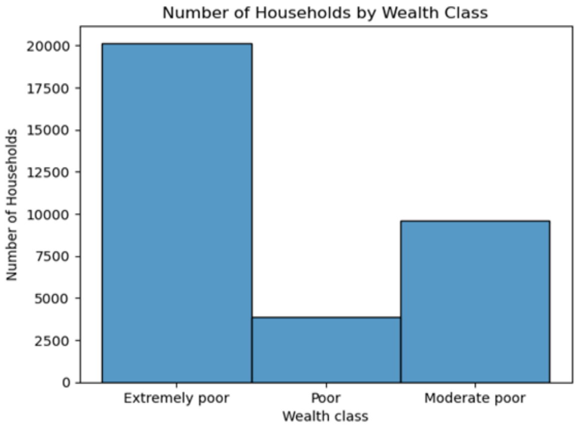


Figure 3

A. Classification Algorithms Considered

1) The Naive Bayes Algorithm

Is another Machine Learning algorithm for classification problems. Naive Bayes is an efficient classification algorithm in data mining that can handle missing values during classification [4] [5]. Naive Bayes Algorithm is pretty fast Machine Learning and efficient model, basically, this model is used for text classification few known examples are spam filtration, sentimental analysis and classifying new articles. The named of Naive is called for it's some sort of features distinct of an event of another feature. And Bayes refers to the statistician and philosopher Thomas Bayes theorem [4]. The NB theorem can be expressed mathematically as follows:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Where:

- P (A | B): Probability of occurrence of event A given the event B is true.
- P (A) and P (B): Probabilities of the occurrence of event A and B respectively.
- P (B | A): Probability of the occurrence of event B given the event A is true.

2) Decision Tree

Is another machine learning classifier that relies on building a tree that represents a decision of instances training data [4]. The Algorithm starts to construct the tree iteratively based on best possible split among features. The selection process of the best features relies on a predefined functions like, entropy, information gain, gain ratio, or gini index.

3) K-Nearest Neighbours

It is one of the simplest Machine Learning algorithms based on Supervised Learning technique. And assumes the similarity between the new case or data and available cases and put the new case into the category that is most similar to the available categories [5]. It stores all the available data and classifies a new data point based on the similarity and easily classified into a well suite category by using K- NN algorithm.

4) Random Forest

Random Forest classifier is a learning method that operates by constructing multiple decision trees and the final decision is made based on the majority of the trees and is chosen by the random forest. It is a tree-shaped diagram used to determine a course of action [5]. Each branch of the tree represents a possible decision, instance, or reaction. Using of Random Forest Algorithm is one of the main advantages is that it reduces the risk of over fitting and the required training time. Additionally, it also offers a high level of accuracy. It runs efficiently in large databases and produces almost accurate predictions by approximating missing data. Using of Random Forest Algorithm is one of the main advantages is that it reduces the risk of over-fitting and the required training time. Additionally, it also offers a high level of accuracy and produces highly accurate predictions by estimating missing data.

5) Support Vector Machine

It is one of the most popularized Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, basically, it is used for Classification problems in Machine Learning scenario [5]. The intent of the SVM algorithm is to create the best decision boundary that can segregate n-dimensional space into classes so that it can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane of SVM.

6) Logistic Regression

Logistic regression is a supervised learning binary classification algorithm. It is used to predict the probability of a target variable. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection, etc. It can also be thought of as the classification version of Linear Regression that uses a special S-shaped curve known as the Sigmoid function.

7) AdaBoost

Ada-boost or Adaptive Boosting is an iterative ensemble boosting classifier. It builds a robust classifier by combining all poor performing classifiers to get the high accuracy, the concept behind Adaboost is to set the multiple weighs of classifiers and train the data in each iteration, hence it ensures the exact prediction of unusual observation [5] [6]. AdaBoost refers to a particular method of training a boosted classifier. Adaboost classifier is a classifier in the form of Where each f_t is a weak learner that takes an object X as input and returns a value indicating the class of the object.

B. Selection of Classification Algorithm

Seven classification algorithms are compared in this study, which are Naïve Bayes, Decision Tree, k-Nearest Neighbors Random Forest, Support Vector Machine, Logistic Regression and AdaBoost. The performance between all the seven classifiers are then evaluated and compared.

DOI: https://dx.doi.org/10.21275/SR25210190332

C. Regularization

Inductive learning is a key concept in machine learning, which refers to the process of learning the general concepts from specific examples provided. Specifically, this refers to the attempt to learn target function from available training data. Meanwhile, **generalization** refers to how well a machine learning model learns the concepts so it can be applied to specific examples that were not seen by the model during the learning.

Machine learning's primary objective is to generalize well enough so that the model obtained from the training set can be applied to the unseen data portion from the problem domain. If the algorithm fits the training data too well, it can cause overfitting issues which may lead to poor performance of a machine learning algorithm [8]. Overfitting is an occurrence where the model learns a both the target function and noise during the training, consequently degrading the performance of that model on an unseen data. Overfitting is more likely to happen to nonlinear models that have more flexibility when learning a target function. A number of methods are available to avoid overfitting. For example, a pruning process can be applied to a decision tree model to reduce the size of a tree that are too large and deep [9], and such process was implemented in this study. Other methods were also applied to overcome overfitting related to specific machine learning models, specifically resampling techniques such as SMOTE oversampling, used to treat imbalanced dataset, and k-fold Cross validation, which iteratively trains and test a model k-times from different training data subsets, to increase generalization chances

4. Literature Review

Poverty has been a decades-long issue which governments, policymakers and organization bodies try to eradicate for years. It is usually determined by the level of income and the sufficient basic resources to maintain sustainable livelihoods. In fact, it also includes the display of education inaccessibility, hunger and malnutrition, social unfairness, and limited access to other basic needs. Having a job does not guarantee that one will have a decent living.

In order to categorise the B40 population in Malaysia, a study was done to examine how well Naive Bayes, Decision Trees, and k-Nearest Neighbours performed [9]. The study made use of the 'eKasih' dataset from the National Poverty Data Bank. It includes a thorough profile of Malaysia's low-income households. The importance of this study emphasises feature engineering, normalisation, sampling technique selection, feature selection approaches, and parameter tweaking.

To balance out the dataset, a technique known as Synthetic Minority Oversampling Technique (SMOTE) is used to construct replica cases. The different combinations of parameters used to optimise each classifier include discretization for Naive Bayes, confidence factor and the minimum number of objects for Decision Trees, as well as k-value and distance function for kNearest Neighbours.

By ranking the top eight features utilising symmetrical uncertainty, correlation, and information gain attributes, feature selection algorithms have been found to increase classification accuracy and the Kappa statistic. K-Fold Cross-Validation of 10 is used as the measurement for assessing the performance of all three classifiers after parameter tuning, and a statistical test is run to determine whether two "models are statistically significantly different from one another or if one of them is better than the other [9].

According to the study's findings, the Decision Tree model is absolutely exceptional and has outperformed other classifiers in terms of accuracy.

Additionally, the factors of poverty at the neighbourhood and individual levels were explored in [10] in a general population of Hong Kong. Prior research mostly concentrated on using financial indicators to evaluate poverty and poverty within a certain population. However, this study places a strong emphasis on a holistic approach to address all aspects of what determines poverty [10]. which uses the deprivation of poor people's experience such as income insufficiency, poor health and education absence.

The author used Quantile Regression to further analyse the differences in the effects of the determinants across five poverty spectrums after using Logistic Regression to study the determinants of poverty. When the poverty line or threshold is employed as the measurement, logistic regression is typically used to assess the level of poverty. Only the percentage of persons who live in poverty may be determined using this method. It does not provide a variety of explanations for the experiences of the impoverished. A more thorough explanation of poverty status is provided by Quantile Regression, which identifies the differential outcomes of the factors that determine poverty across the poverty spectrum. Based on the ratio of income to poverty (I/P), quantile regression calculates the level of poverty. This study defines six poverty thresholds based on the poverty line for Hong Kong in 2015. These thresholds apply to households with up to six people.

Additionally, a specific amount of I/P ratio is mapped to five quantiles. Each quantile is mapped to a particular category of poverty status. The quantile regression model's findings provide magnitudes of connections between different factors and poverty status, including whether a given variable is significantly or not across the range of poverty and whether it is positively or negatively linked with poverty. Ordinary Least Square (OLS) regression is used in statistical analysis for purposes of comparison of "how some quantiles of the I/P ratio may be more affected by a certain predictor than other quantiles" [10].

The use of cluster-robust standard errors and data weighting for the oversampled data are also described. The poverty prediction task was investigated in [10] based on household income using satellite images in the urban environment of North and South America. The cost and labor-intensiveness of the approach for gathering socioeconomic data served as the study's driving forces [10]. As a result, it turns to remote sensing data, which is more suited for estimating poverty on a big scale, including high-resolution satellite photography. To create a descriptive urban landscape for identifying the urban areas in each city, original satellite data is combined with Open Street Map (OSM) crowdsourced data.

Regression and convolutional feature extraction are used in the study to estimate the location of objects. Transfer learning using three ConvNets: ResNet50 with initialised ImageNet weights, VGGF trained on ImageNet weights, and VGGF fine-tuned with nightlight intensity from a few African nations. (Jean *et al.*, 2016) are used for features extraction. Two separate tiers of census area boundaries are used to extract two different sorts of socioeconomic information from an input image. Before mapping, each input image is additionally rotated and flipped either horizontally or vertically. ResNet50 feature is picked as the model after all three neural networks have undergone cross-validation. In the meantime, the Ridge Regression model is used to complete the family income prediction task based on picture level characteristics and cluster level information. Regression score, metrics score, and 10-fold cross-validation are used to assess the model's performance.

To discover the elements that define the condition of poverty, ordinal and multinomial logistic regression models were examined [11]. According to this study, there are three levels of poverty: absolute poverty, near poverty, and above near poverty. Based on the household income percentage below the poverty criterion, which is set at 100% to 125% for near-poverty states and greater than 125% for states that are above near-poverty states, each state is evaluated. The threshold is multiplied by the inflation rate for each succeeding year and is based on the national median income for Poland in 2000. Additionally, the information is derived from multiplying the number of households between 2000 and 2015 every two years.

According to [11] In the Social Diagnosis 2015 study, two questionnaires were used to gather information from families. The first survey involves face-to-face interviews with the household substitute, who is the expert on the members of the home and their current situation, to gather information about the make-up of the household and living conditions. It provides a wealth of information on household composition, living conditions, and the demographic and socioeconomic circumstances of each household member.

All household members who are 16 years of age or older are asked to complete the second survey. It focuses on issues that reveal a person's level of well-being. Gender, age, education level, residence, number of household members, biological family type (e.g., single without children, married couple with children, etc.), socioeconomic group, employment status, and presence of a disabled person in the household are among the variables analysed to [11]. According to the findings, the multinomial logit model performs better in predicting the level of poverty. Because the ordinal logistic regression model does not satisfy the requirement of parallel lines, the results may be misinterpreted. The study also notes that the variables (education, place of residence, labour force participation, and socioeconomic category) are the most important influences on the state of poverty [11].

The authors in [12]. assessed the factors that matter to poverty by predicting the poverty level through a machine learning approach. They used data from two sources: The Oxford Poverty & Human Development Initiative, which offers the poverty index for various nations, and the Poverty Possibility Index, which contains information about individuals. The data analysis was carried out by the authors to get insights into and establish relationships between several factors that may influence the likelihood of poverty. The studies investigated various machine learning techniques for the best model for predicting and categorising poverty, including linear regression, decision trees, random forests, gradient boosting, and even neural networks. According to the study's important score, the variables that contribute to poverty are highlighted.

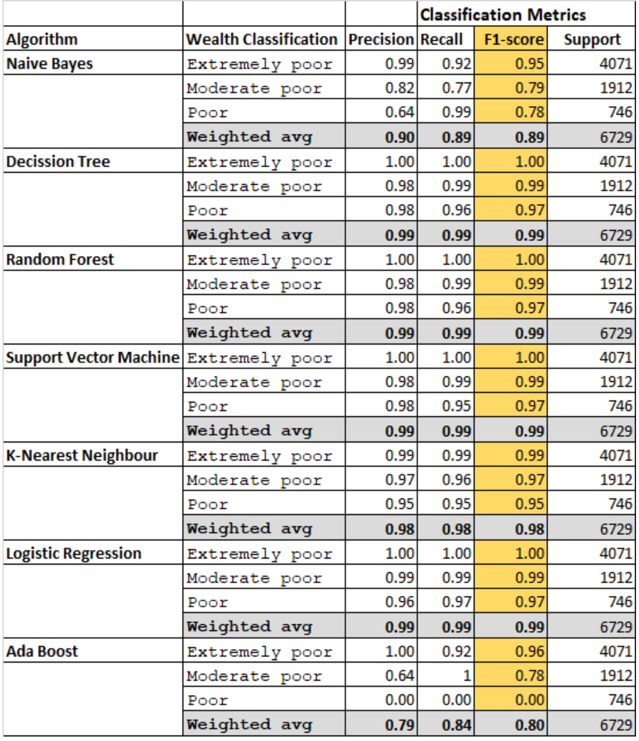
The authors came to the conclusion that, in terms of accuracy, dependability, and complexity, gradient boosting classifier is superior to other methods. The country they live in is the second biggest contributor to poverty after the amount of education. In contrast to [12] that examines the large population across countries, the authors in [13], [14] and [15] explored specific countries for poverty study, like Jordan and China, respectively. Another study [13 focused on the poverty problem in Jordan based on five years (2002, 2006, 2008, 2010, 2017) survey data on income and household expenditures. The research utilizes the machine learning framework as a simple, inexpensive and accurate tool.

Since the acquired data is substantially unbalanced, this study emphasises data pre-processing and imbalanced data management as its primary contribution. For two target classes (poor and non-poor), random oversampling, class weight, and SMOTE (Synthetic Minority Oversampling Technique) are used to address this issue. The 16 various machine learning models are trained to pick the top two models for determining poverty in real time.

Out of sixteen models, the LightBGM and Begged Decision Trees have the highest accuracy. The research was conducted in (Liu *et al.*, 2021) to identify the poverty determinants in a rural area of China, Yunyang County. In contrast to previous research, which has mostly focused on predicting poverty, this study explored the significance of various feature variables and how they relate to poverty at the village level. The study incorporated the survey data of people's information gathered in 2017. The authors used the Lindeman, Merenda, and Gold (LMG) approach in multiple linear regression and the random forest classifier (RF) to determine the critical variables influencing the distribution of poverty at the village level.

5. Results and Discussions

Table 3: Models Classification Report Results



		Classification Metrics			
Algorithm	**Wealth Classification**	**Precision**	**Recall**	**F1-score**	**Support**
Naive Bayes	Extremely poor	0.99	0.92	0.95	4071
	Moderate poor	0.82	0.77	0.79	1912
	Poor	0.64	0.99	0.78	746
	Weighted avg	**0.90**	**0.89**	**0.89**	6729
Decission Tree	Extremely poor	1.00	1.00	1.00	4071
	Moderate poor	0.98	0.99	0.99	1912
	Poor	0.98	0.96	0.97	746
	Weighted avg	**0.99**	**0.99**	**0.99**	6729
Random Forest	Extremely poor	1.00	1.00	1.00	4071
	Moderate poor	0.98	0.99	0.99	1912
	Poor	0.98	0.96	0.97	746
	Weighted avg	**0.99**	**0.99**	**0.99**	6729
Support Vector Machine	Extremely poor	1.00	1.00	1.00	4071
	Moderate poor	0.98	0.99	0.99	1912
	Poor	0.98	0.95	0.97	746
	Weighted avg	**0.99**	**0.99**	**0.99**	6729
K-Nearest Neighbour	Extremely poor	0.99	0.99	0.99	4071
	Moderate poor	0.97	0.96	0.97	1912
	Poor	0.95	0.95	0.95	746
	Weighted avg	**0.98**	**0.98**	**0.98**	6729
Logistic Regression	Extremely poor	1.00	1.00	1.00	4071
	Moderate poor	0.99	0.99	0.99	1912
	Poor	0.96	0.97	0.97	746
	Weighted avg	**0.99**	**0.99**	**0.99**	6729
Ada Boost	Extremely poor	1.00	0.92	0.96	4071
	Moderate poor	0.64	1	0.78	1912
	Poor	0.00	0.00	0.00	746
	Weighted avg	**0.79**	**0.84**	**0.80**	6729

As shown by the results above I used weighted F1-score as the model performance metrics this is because the dataset distribution of wealth class was not evenly distributed. The weighted F1-score Logistic regression, Support Vector Machine, Random Forest and Decision Tree they were all 0.99 weighted F1-score. During feature selection, tree-based technique was applied for selecting best features, by so doing tree-based algorithms are not considered for this model. This means even though Random Forest and Decision Tree had weighted F1-score of 0.99 they are not considered for this model. Comparing precision and recall of Logistic regression, Support Vector Machine, Logistic regression had high weighted average scores for precision and recall for all the three wealth class labels. K-Fold cross validation technique, was applied to Logistic regression the weighted average F1-score remains at 0.99. Stratified K-Fold of 10 folds was calculated on the whole dataset and average accuracy was calculated. Logistic Regression was considered for this model due to its high performance score on weighted average F1-score.

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

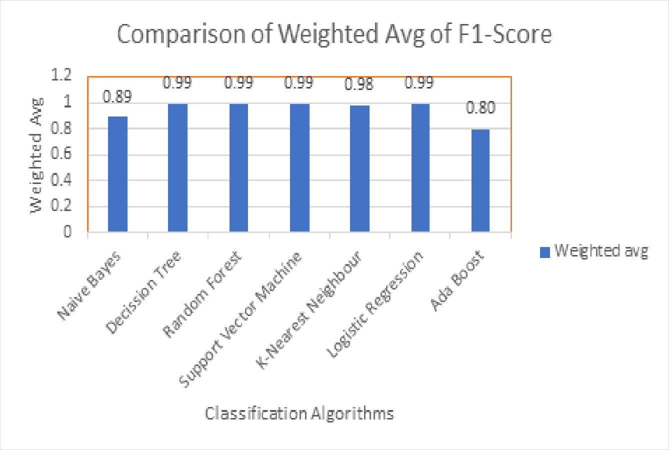


Figure 4: Comparison of Weighted Average of F1-Score

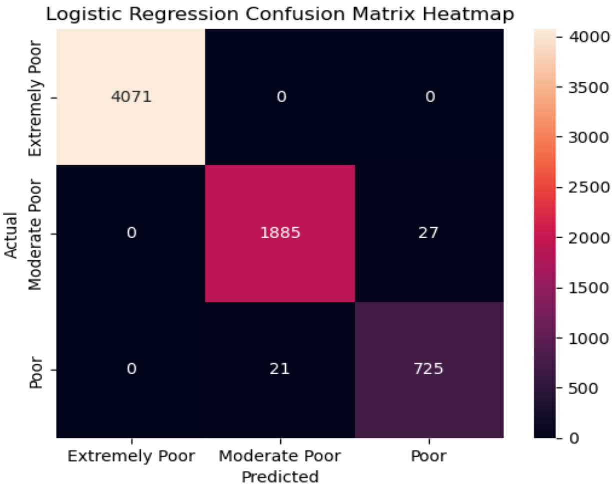
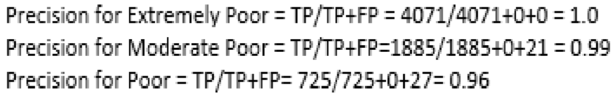


Figure 5: Logistic Regression Confusion Matrix Heatmap

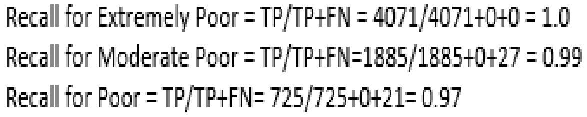
Calculations of Model Performance Evaluation Metrics

Precision



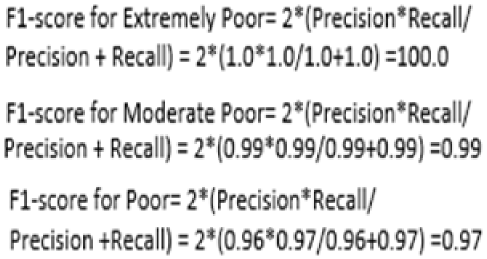
Precision for Extremely Poor = TP/TP+FP = 4071/4071+0+0 = 1.0
Precision for Moderate Poor = TP/TP+FP=1885/1885+0+21 = 0.99
Precision for Poor = TP/TP+FP= 725/725+0+27= 0.96

Recall



Recall for Extremely Poor = TP/TP+FN = 4071/4071+0+0 = 1.0
Recall for Moderate Poor = TP/TP+FN=1885/1885+0+27 = 0.99
Recall for Poor = TP/TP+FN= 725/725+0+21= 0.97

F1-Score

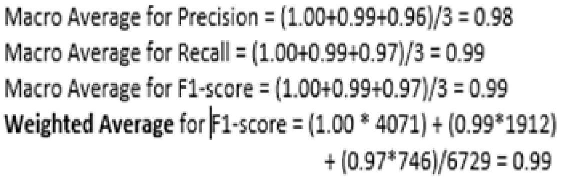


F1-score for Extremely Poor= 2*(Precision*Recall/
Precision + Recall) = 2*(1.0*1.0/1.0+1.0) =100.0

F1-score for Moderate Poor= 2*(Precision*Recall/
Precision + Recall) = 2*(0.99*0.99/0.99+0.99) =0.99

F1-score for Poor= 2*(Precision*Recall/
Precision +Recall) = 2*(0.96*0.97/0.96+0.97) =0.97

Macro Average and Weighted Average



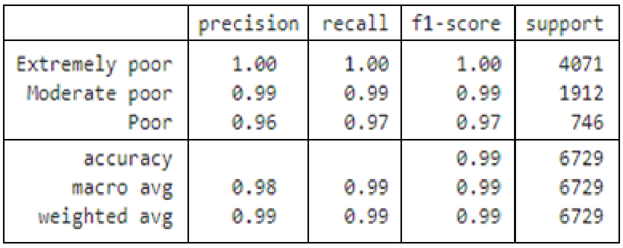
Macro Average for Precision = (1.00+0.99+0.96)/3 = 0.98

Macro Average for Recall = (1.00+0.99+0.97)/3 = 0.99

Macro Average for F1-score = (1.00+0.99+0.97)/3 = 0.99

Weighted Average for F1-score = (1.00 * 4071) + (0.99*1912) + (0.97*746)/6729 = 0.99

Logistic Regression Classification Report



	precision	recall	f1-score	support
Extremely poor	1.00	1.00	1.00	4071
Moderate poor	0.99	0.99	0.99	1912
Poor	0.96	0.97	0.97	746
accuracy			0.99	6729
macro avg	0.98	0.99	0.99	6729
weighted avg	0.99	0.99	0.99	6729

6. Conclusions and Further Work

With the increase use of machine learning in different domains, it has become a requirement to develop machine learning web-based system to predict households that are in different levels of poverty. The stimulus for the study was to identify and implement the best and efficient machine learning algorithm for classifying rural households according to their level of poverty. This study identified Logistic Regression algorithm as the best algorithm for predicting poverty-stricken households. In the future, the study can be extended to cover households in urban areas. Indicators or features considered for classifying rural households should be continuously revised so as to be in line with ever changing poverty definition.

References

[1] Min, P.P. *et al.* (2022) 'Poverty prediction using machine learning approach', *Journal of Southwest Jiaotong University*, 57(1).

[2] Alsharkawi, A. *et al.* (2021b) 'Poverty classification using machine learning: The case of Jordan', *Sustainability*, 13(3), p. 1412.

[3] Hu, L. Y., Huang, M. W., Ke, S. W., & Tsai, C. F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus, 5(1), 1304.

[4] Shreem, S. S., Abdullah, S., & Nazri, M. Z. A. (2016). Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm. *International Journal of Systems Science*, 47(6), 1312-1329.

[5] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16- 28.

[6] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929-1958.

[7] Sani, N.S. *et al.* (2018) 'Machine learning approach for bottom 40 percent households (B40) poverty classification', *Int. J. Adv. Sci. Eng. Inf. Technol*, 8(4–2)

[8] Alhutaish, R., & Omar, N. (2017). Feature Selection for Multi-label Document Based on Wrapper Approach through Class Association Rules. *International Journal on Advanced Science, Engineering and Information Technology*, 7(2), 642-649

[9] Samsiah Sani, N., Shlash, I., Hassan, M., Hadi, A., & Aliff, M. (2017). Enhancing Malaysia Rainfall Prediction Using Classification Techniques. *J. Appl. Environ. Biol. Sci*, 7(2S), 20-29.

[10] Peng, C. *et al.* (2019) 'Determinants of poverty and their variation across the poverty spectrum: Evidence from Hong Kong, a high-income society with a high poverty level', *Social Indicators Research*, 144, pp. 219–250.

[11] Sączewska-Piotrowska, A. (2018) 'Determinants of the state of poverty using logistic regression'.

[12] Zixi, H. (2021) 'Poverty Prediction Through Machine Learning', in 2021 2nd *International Conference on E-Commerce and Internet Technology (ECIT)*. IEEE, pp. 314–324.

[13] Alsharkawi, A. *et al.* (2021a) 'Poverty classification using machine learning: The case of Jordan', *Sustainability*, 13(3), p. 1412.

[14] Jean, N. *et al.* (2016) 'Combining satellite imagery and machine learning to predict poverty', *Science*, 353(6301), pp. 790–794.

[15] Liu, M. *et al.* (2021) 'Using multiple linear regression and random forests to identify spatial poverty determinants in rural China', *Spatial Statistics*, 42, p. 100461.