National College of Ireland

Poverty Prediction of Developing Countries Based on Household Survey Data using Ensemble Methods

Research in Computing - CA 2
MSc in Data Analytics

Heena Chopra
Student ID: x19205309

School of Computing
National College of Ireland

Submitted to : Prof. Sean Heeney

Contents

Poverty Prediction of Developing Countries Based on Household Survey Data using Ensemble Methods

Heena Chopra
x19205309

Abstract

Measuring poverty is a complex yet an important aspect in studying the economical and social growth of any developing country. Throughout the years, there have been many different approaches based on many different kinds of data to precisely predict the poverty and poverty based indices. The measurement of poverty is an extremely time consuming, yet crucial part of the economy of a country. It has a been estimated that the COVID-19 pandemic has pushed around 90 million households into extreme poverty in the year 2020. In this study, we will propose a novel technique based on ensemble learning to predict the poverty for a given household on the basis of the survey data collected from people. The data has been made available by the world bank and is open source. The data contains the information about household and individuals for three different countries which drives the choice of different combinations of ensembling algorithms. We propose three different architectures for the tree countries based on the ensemble of number of neural networks and LightGBM algorithms. We also propose a 20-fold cross validation procedure to quantify the relationships between different features. The results will be evaluated on the basis of the mean log loss for each country.

1 Introduction

Most of the people on this planet are living in extreme poverty. According to the latest statistics, 85% of the world live on less than $30 per day, two-thirds live on less than $10 per day, and every tenth person lives on less than $1.90 per day (*Poverty* (2021)). It is said that even today every tenth person in this world is living in extreme poverty. The rise in world population growth has caused an extreme shoot-up in the number of poor people across the globe. It has become very important to measure poverty because this trend of increasing poverty strikes fears in the heart of the upcoming decade. Despite this fact, it is extremely hard, difficult, and expensive to classify and predict poverty. Measurement of poverty has two major impediments. (1)- Poverty Identification, (2)- Creation of an index to measure poverty. Generally, Income calculation is used to solve the first problem but the second one is a big topic of discussion amongst all the researchers and practitioners. Researchers have brought up many ways to handle the second complication and proposed novel poverty measurement indices, which are later described in detail in the section (2).

In the last years, a huge amount of work has been done on poverty prediction. Several studies highlight the difference in the accuracy of the results when compared between single and multiple data sources (Pokhriyal and Jacques (2017); Blumenstock et al. (2015)). The above studies prove that there are many ways for poverty estimation and

measurement i.e.; income-based, nutrition-based, consumption-based, etc. Proxy means test (PMT) has become a common tool for poverty measurement from the visible characteristics of any household when the wealth or income statistics are not present (Grosh and Baker (1995)). Economic changes and the household surveys happen to create a gap after some time when the household survey data gets outdated by the new economic changes, this is when Call detail records (CDR's) can help in predicting welfare-based on-call activities. (Hernandez et al. (2017)). Satellite imageries have brought a novelty in the poverty identification of rural and urban areas of many countries (Jean et al. (2016); Subash et al. (2018)). The satellite imageries being a single source of data set back a limitation which can be resolved by combining the satellite imagery data with geospatial data (Tingzon et al. (2019)). Although, using satellite images has produced state-of-the-art results in the domain of poverty prediction but the use of household survey data provides better predictors for determining poverty.

The World Bank aims to end poverty by 2030. In this study, we predict whether a household is poor or not based on household survey data from three different developing countries, each having its own demographic features. With the same background and iterating to all the studies of feature selection, this study expands the approach through ensembled models i.e., Light GBM (Gradient Boosting) and Neural Networks. This is an accurate, inexpensive, and scalable method for poverty identification, classification, and measurement. The model implementation and evaluation for the study is explained are detail in sections 5 and 6.

1.1 Background and Motivation

Poverty is likely to rise in the coming years and elimination of poverty until the year 2030 is the topmost Sustainable Development goal. The extent of poverty is the biggest subject of importance today. The number of people below the poverty line has increased suddenly due to the COVID-19 Pandemic i.e.; COVID-19 has induced new poor. During this pandemic, jobs have vanished at a startling rate creating high unemployment. This pandemic is responsible for bringing 71 to 100 million people below the international poverty line and pushing them towards extreme poverty, in the year 2020. More than 1 billion people have fallen below the international poverty line and a major fraction of these numbers can be attributed to the COVID-19 Pandemic (1). World Bank aims to revise the number to end this extreme poverty, keeping in mind the pre-COVID-19 and post-COVID-poor and non-poor count.

A lot of work has been done on Poverty estimation and measurement. Each study highlights a novel way of classifying and measuring poverty. Generally, poverty has been measured using survey data of wealth and income of households (Blumenstock (2016)). Later, many researchers decided to measure poverty using remote sensing techniques, like daylight and nightlight imagery. They explored the potential of using satellite imagery in measuring poverty. (Jean et al. (2016)). In few types of research, Nightlight imagery data was combined with daytime satellite imagery to obtain better accuracy (Ni et al. (2020)). From many studies, it is clearly explained how social media and internet data can also be used as a proxy for the economic activities data to measure poverty. (Engelmann et al. (2018)). In our research, we have iterated the previous studies and developed a novel methodology for poverty measurement in developing countries using household data.

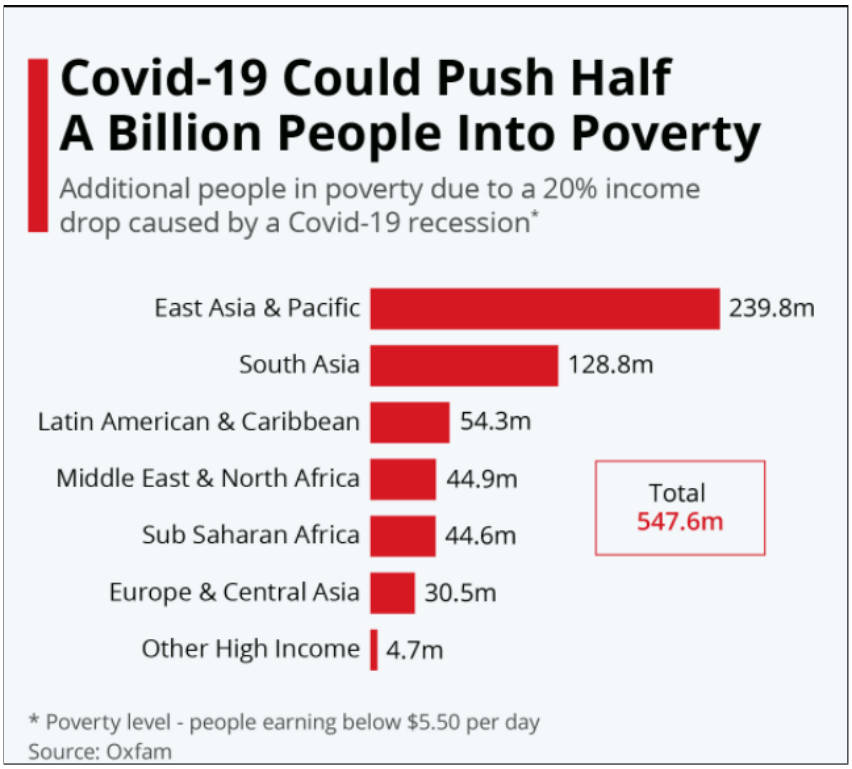

Figure 1: Increasing poverty trends due to COVID-19 pandemic

1.2 Research Question

How can the different ensemble learning algorithms employing parameter optimization and feature selection techniques be applied in the prediction of poverty of a given household of a country based on its survey data?

The main objective of this study is the precise prediction of poverty for a given household i.e classify a household as poor or not poor, given it's survey data by employing ensemble learning technique. The methods to be ensembled together are the conventional neural networks and a gradient boosting algorithm known as LightGBM. The choice for the ensemble algorithms was made after critically reviewing the existing literature in the domain. Moreover, it has been observed that using ensemble methods boosts the overall prediction accuracy for the problem.

1.3 Research Objectives

The following are the research objectives for our study:

- To perform feature engineering and optimize the different parameters for the given household survey data.
- To apply an ensemble of machine learning algorithms for the poverty prediction for a given household and evaluate the performance using cross-validation.

2 Literature Review

2.1 Introduction

The literature review has been divided into the following two parts based on the previous works done in the domain:

- **Poverty Prediction using Feature Based Data**: This section consists of the literature which employs feature based data for the prediction of poverty.
- **Poverty Prediction using Image Based Data**: This section consists of the literature which employs data based on the remote sensing images (from a satellite) for the prediction of poverty.

2.2 Poverty Prediction using Feature Based Data

The world economy is growing very fast. For economic research, it is very important yet difficult to get accurate and timely statistics of population characteristics. In small areas, the major challenge to researchers of the public and private sector is the lack of reliable quantitative data as they require data in disaggregated form i.e., data broken down into detailed subcategories like age, gender, and ethnicity, which often do not exist (Elbers et al. (2003), Omrani et al. (2009)). Therefore, novel sources of data are being explored for better modelling and measurement like social media have been used for the measurement of economic development (Eagle et al. (2010)), unemployment (Choi and Varian (2012)), and electoral outcomes (Wang et al. (2015)). Similarly, Blumenstock et al. (2015) introduced the use of mobile phone data for poverty prediction, where they show how an individual's digital footsteps that is the history of mobile phone use can be used in predicting his/her socioeconomic characteristics. They further explained that the overall wealth or asset distribution patterns across the nation or sub-regions can be reconstructed as per the predicted values of individual's characteristics. The future integrations to this approach can be to remove the limit that the model is overfitted on a single data source. Later, in an interesting research done by Pokhriyal and Jacques (2017), a novel machine learning technique was used to predict poverty. Their approach was different from the prior works that have examined poverty measures using a single data source, and they focused on how to get multiple datasets together in poverty prediction. In this, a Gaussian Process regression (GPR) computational framework was designed to predict the multidimensional poverty index (MPI) using environmental data and call data records (CDR's). Gaussian Process regression (GPR) is a non-parametric and kernel-based algorithm i.e., a class of algorithm that deals with pattern analysis (Task of finding and studying different types of relationships between various datasets) (Rasmussen (2003)). This technique is quite useful when we integrate disparate data sources as this method doesn't require different data ecosystems to share the data between them and therefore the privacy for all the datasets keeps maintained. Therefore, the computational framework came out with one of the best accurate results i.e., a Pearson correlation of 0.91. They also mentioned that relying on CDR's for poverty measurement can be a major issue at one point of time because CDR's are maintained by different private service providers of the country and unless all the providers share the data, the estimation is likely to be incomplete and therefore more methods for poverty measurement can be explored. Integrating two data sources has shown commendable results and therefore multiple source

big data has become a new approach to measure poverty. Niu et al. (2020) integrated social media data with remote sensing images to form a Multi-source Data Poverty Index (MDPI) combined with random forest machine learning algorithm to measure poverty characteristics.

The mission at the end for any organization is to serve the poor and end poverty. However, to complete this mission, they need to identify the poor households but this could be quite expensive, time-consuming, and difficult for data-scarce countries due to lack of information. The best alternative to this is the Proxy means test (PMT), first developed during the 1890s to target the social programs in Latin America (Caldés et al. (2006)). Proxy means test is a methodology in which we predict whether a household is poor or not using only a small amount of data, perhaps a small questionnaire of 10 -30 questions from nationally available household survey data (Schreiner (2010)). In the Proxy means test, a subset of the attributes is chosen from the survey and then a model is applied using these attributes followed by testing of the results using held-out data to predict poverty at the household level (Kshirsagar et al. (2017)) constructed a new methodology for Proxy means test i.e., the Poverty Probability Index (PPI) using conventional cross-validation and parameter regularization techniques. Instead of spending several hours in getting the poverty measures for households, Poverty Probability Index (PPI) helped them get 10 small questions that can be used within all subregions and all poverty lines of these diverse countries. They limited their study to the use of bootstrap variables, probability, and statistics in measuring poverty and did not involve any interaction between their predictors. Similar to them were McBride and Nichols (2015). who agreed upon the fact that PMT's can help in making poverty assessment much cheaper as compared to a full means test that is costlier. In this paper, they presented how the implementation of machine learning algorithms to Proxy means test shoot up the overall performance of the target. They worked upon the United States Agency for International Development (USAID) poverty assessment tools and base data for modelling, method implementation, and demonstrations. Classification and regression tree methodology, also known as the CART was introduced in 1984 by Leo Breiman , Jerome Friedman, Richard Olshen and Charles stone(medium url). While both of the tree methodologies have their own algorithms to achieve the target variable, they focused on regression trees and, in particular quantile regression trees in their study as it is best suited for prediction like problems. The quantile regression method showed a rise in out-of-sample performance from 2 to 18 percent as compared to the previous methods. They later suggested exploration of more machine learning algorithms for better results. Unlike the current common practices of estimating poverty using different variations of linear regression, Fox and Sohnesen (2016) evaluated the Random Forest method for poverty prediction. Random Forest being one of the best classifiers ensembled model, consists of a large number of individual decision trees that have their own class predictions and the class with the greatest number of votes becomes the model's final prediction. Random forests outperform the decision trees generally, and also are less overfitted to their respective training sets as compared to other decision trees (Breiman (2001)). In this study, they analysed the Random Forest's predictive performance at the rural and urban levels of Ethiopia and found out that Random Forest has a higher accuracy as compared to the existing works done on Regression-based models. Random Forest is a very simple and automated machine learning model, that could be used instead of other models. They also mentioned that the accuracy of Random Forest is not always that accurate, for example predictions at a national level. The difference in accuracy is not significant, but this highlighted

the limit that more machine learning algorithms options could be explored for better accuracy results. Within the recent literature for poverty predictions, one of the studies with application of Random forest was done by Ruben Thoplan, who addressed the issue of poverty in Mauritius. He used random forest classifier for categorising the people of Mauritius basis the relative poverty line i.e.; distinguishing the poor from the non-poor people. Random Forests has a number of benefits over the other decision trees, but one of the best features of this algorithm is that it can rank the variables in order of their importance. In this study, the main classifiers for the poverty were calculated to be the number of hours worked per week, age, education, and sex of any individual of the country. He had two main observations after modelling and method implementation. First, Random forest is one of the best data mining classification tools for poverty measurement as the out-of-bag error was very low. Second, there is huge gender-poverty void in Mauritius as the female were the most classified poor as compared to men who were highly classified as non-poor. This study integrated to the existing methods of poverty alleviation, by simply classifying the variables that top the charts in poverty prediction, targeted to which policy and opportunities related decisions could be made for just ending poverty.

There are two main approaches for machine learning; Supervised machine learning, a task driven approach where we have a labelled dataset for training the algorithm and Unsupervised machine learning where the model is handed over a dataset without any explicit instructions (Ozgur (2004)). Supervised machine learning approach is applied on classification problems, where data is pre – categorized. This approach has been used by many researchers in poverty measurement as well, just like Sani et al. (2018). Al who applied one of the best machine learning models i.e.; Naive Bayes, Decision Tree and k-Nearest Neighbors for classifying the Bottom 40 (B40) population of Malaysia. Decision tree model outperformed the other models and gave one of the most significant results. Similarly, in one of the studies by Alsharkawi et al. (2021) various classification algorithms were applied but LightGMB achieved one of the best performances with 81% F1-Score. They measured the multidimensional poverty problem in Jordanian households using the household expenditures and income surveys. Feature's selection is a very important aspect for classification of each class of poverty. (Mohamud and Gerek (2019)). In their study, they considered various household characteristics for poverty identification and mined the right features that emphasize the poverty. Considering Feature Selection, one of the most important aspect of poverty identification and measurement Ferreira et al. (2020) they combined supervised and unsupervised machine learning algorithms for achieve their research objectives.

2.3 Poverty Prediction using Image Based Data

Data on key features of the economic development are still unavailable and this data gap for any continent or sub-continent is very constraining. An alternative path to this method of collecting data, is to use novel sources of data such as Satellite imageries. This technique has shown promising results earlier in improving the economic production statistics. After reviewing the reliable results of the technique in other socio-economic related issues, this technique was used for poverty prediction. Night light imagery has given promising signs of providing accurate and up to date indications of poverty. (Henderson et al. (2012); Michalopoulos and Papaioannou (2013); Pinkovskiy and Sala-i Martin (2016)). Taking Nightlight Satellite imagery as their starting point, authors have begun

to predict economic activity indicators such as poverty. High frequency night light data, trained using Artificial Neural Networks turned out to be a better predictor than Per Capita Income at state i.e.; sub national level (Subash et al. (2018)). Artificial Neural Networks have been used for modeling and prediction by many other researcher's as well because of its promising results. Nischal et al. (2015) reported that night light satellite imagery can successfully estimate the poverty density level for any region. However, these studies somehow are restricted to statistical data as the researchers combined the satellite imagery data with statistical data to measure poverty. Also, they focused their studies to single features such as distribution features of night light, central tendency etc. Unlike before, Li et al. (2019) proposed a study where they demonstrated how high poverty data can be estimated and measured by machine learning algorithms using only Defense Meteorological Satellite Program/Operational Linescan System (DMSP/OLS) night light imagery data. Nightlight imagery data provides abundant information that can be related to poverty but it may be insufficient to just rely on nightlight satellite data. Introducing, a novel methodology for data identification and estimation Yin et al. (2021) carried-out poverty identification and measurement in Guizhou Province, China, 2012 using night light satellite data and geographical data. They extracted 23 spatial features from the integrated dataset using random forest machine learning algorithm. Similarly, Masoomali et al. (2020) mapped the distribution of poverty in Philippine archipelago using geographic information from OpenStreetMap (OSM) combined with Nightlight satellite data. They observed a 63% variation in the results when compared to previous studies that just used nightlight satellite data as the data source. It is appreciable to use high-resolution satellite imageries for poverty prediction but at the same time it is important to highlight the fact that these imageries are extremely expensive to purchase and comes at a high cost. In one of the studies, Ayush et al. (2020) reduced the purchase cost of the imageries and to maintain accuracy, they proposed a novel reinforcement learning technique where they first let free low-resolution imagery identify the places where costly high-resolution imageries are highly acquired. In this approach they've created a reward function on two of the real time constraints i.e.; budget and GPU availability and trained a network to approximate the object counts for these constraints in a particular location. They showed that this approach has enhanced the previous performances as they used almost 80% fewer high-resolution images and even the predictive performance was improved prior to previous approaches where all of the high-resolution imageries were used for a particular location. Hersh et al. (2020) Al used Sentinel-2 and MODIS satellites for imagery data, they provide free imagery, covering the area globally. They recorded 8% improvement in the model performance after including these satellite features, as compared to previous models.

Daytime satellite imageries are taken at a much higher resolution as compared to night-time satellite imageries. They are taken in such a way that paved roads, metal roofs all are visible and therefore, it is easier to differentiate between poor and ultra-poor regions. Jean et al. (2016) introduced this novel technique of daytime satellite imagery for poverty measurement and prediction in 2016, using transfer learning approach. Non availability of time series data in developing countries is complication for poverty devising, measurement and implementation leading to which a proxy (noisy but easily obtained) is used to train the deep learning model. Convolutional neural network (CNN) was applied on the publicly available survey and satellite data of five major African countries i.e.; Nigeria, Tanzania, Uganda, Malawi, and Rwanda. They measured the average household consumption and asset wealth across the multiple African countries at a cluster level. The

applied transfer learning models turned out to be strongly predictable. Separate models were trained for each of the African countries where each model showcased a variation of 37% to 55% in the average household consumption. They achieved highly predictive results despite of having a limited training data and lack of temporal labels (The time for each of the imagery was unknown) for the satellite imageries. A similar approach was applied by Piaggesi et al. (2019) upon the poverty metrices of five different metropolitan areas of north and south America using satellite imagery basis pre-trained convolutional models and transfer learning that have been used for poverty prediction in developing areas. Convolutional Neural networks have delivered one of most promising results in this domain, that could be signified from a recent paper by (Perez et al. (2019)) where Convolutional Neural Networks were trained from Landsat 7(NASA (2021)) to build a model on free and publicly available daytime satellite imagery data of African Continent. They found that out of all the satellite imageries only 5% of them were labelled and hence they followed a semi-supervised approach using Wasserstein GAN (Generative Adversarial Networks) regularised with Gradient. Convolutional neural networks have shown promising results, they can be trained end to end on high and medium resolution satellite imageries to measure poverty (Babenko et al. (2017)). In another study, Yeh et al. (2020) considered that it may be possible to measure the poverty of households using satellite imagery which has already been done in many of the previous works but these high-resolution imageries can be quite expensive, not updated that frequently and is generally available only for recent years. They trained Convolutional Neural network (CNN) models on publicly free available daytime satellite imagery data of the African Continent taken from Landsat 7 satellite. This satellite collects the imageries for nearly two decades with a global coverage. Their reported results showed that their current method can only predict when ground data is available for a single country and hence, they found that a combination with nightlights imagery may lead to a further scope of improvement. The accuracy of the results achieved after applying Satellite imageries with a combination of Computational Neural networks in continents like Africa and America, was commendable and needs to be further integrated in more parts of the world and therefore, later this novel technique was proposed by Pandey et al. (2018) to measure poverty in rural areas/regions of India, with an objective of automatically absorbing the features through satellite images that are indicative of poverty. They introduced a two-step approach in which first, a multi task fully convolutional deep network is introduced , for predicting the three main development parameters for any region i.e.; the material of roof, source of lighting and source of drinking water from satellite images and then used these estimated statistics as an input to train another model that predicts the income (direct indicator of poverty). The models that utilized all the three development parameters performed better as compared to the models that utilized only single parameters. Today, it's nearly impossible to have timely and accurate wealth data to predict poverty and to fill this information gap many studies have been done using Satellite imageries combined with machine learning approaches, but Kondmann and Zhu (2020) brought up a novel issue that if this approach can also monitor changes in local poverty over time. They tested this approach in Rwanda from 2005 to 2015 and found out that it struggles to find changes in development from one to another period of time. No changes were captured or predicted, even though Rwanda has witnessed large reductions in local poverty and a significant gap in wealth levels.

Further, many studies investigated the differences in model performance, after combining daytime and night time satellite imageries for poverty measurement. Ni et al.

(2020) proposed a novel method in their study where they have considered a combination of daytime and night-time satellite imagery unlike the previous works that were done purely either on day light or night light imagery. In this proposed method, they involved examination of four deep learning techniques i.e., VGG-Net, Inception-Net, ResNet, and DenseNet, to extract deep attributes from daylight satellite imageries. Further, they built a least absolute shrinkage and selection operator (LASSO) regression model on the deep learning approaches. They later enhanced the performance of ResNet and DenseNet by integrating the Squeeze and excitation (SE) module and focal loss and DenseNet with SE module gave one of the best performances and also validated the importance of SE and focal loss enhancement. A pictorial representation of their methodology is shown in the figure 2

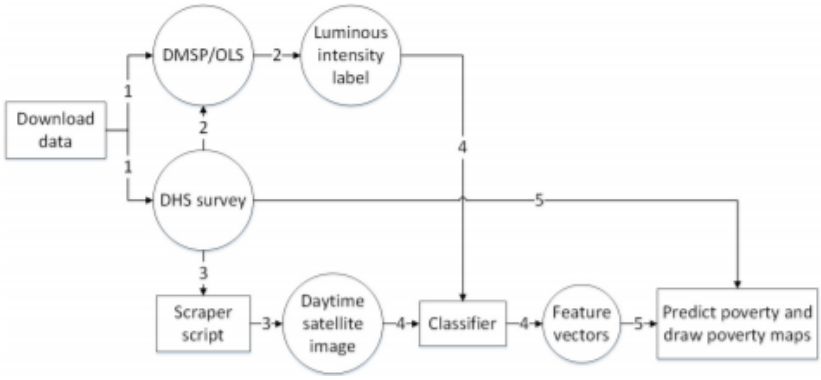

Figure 2: Methodology used by Ni et al. (2020)

2.4 Conclusion

The accuracy results from machine learning approaches have demonstrated how powerful and promising machine learning approaches can be in improving the current country-level economic statistics, but this approach is less capable of finding differences in poverty at the areas with population that lives below or near the international poverty line ($1.90 per capita per day). In these areas, night light imageries are less useful for studying and tracking the households and livelihoods of poor and very poor because of low luminosity levels. Other approaches, such as using mobile phone or CDR's data to measure poverty show promising results but, are difficult to track and trace, given their reliance on service providers. Remote sensing datasets are unique, sensitive and consistent and valuable as compared to the traditional datasets. Above all, it can be said that daytime satellite imagery has delivered much more granular estimates for poverty predictions in urban and rural neighbourhoods' and when combined with night light satellite imagery, gives commendable and promising results. In this study, we are using the household level survey data to predict poverty using an ensemble of machine learning methods. The ensemble we are using consists of combination of different no. of neural networks and Light gradient boosting (LightGBM) algorithms. The reason for the choice of ensemble methods can be justified by the fact that, when combined with a gradient boosting algorithm like XGBoost or LightGBM, the classification results obtained are optimized.

3 Methodology

Methodology is one of the most crucial elements of any research proposal, in which we analyze theoretical approaches and methods that are going to be employed and applied in the research. A well-defined methodology comforts us in achieving high accuracy, low cost and time management. For this study, Knowledge discovery in Databases (KDD) methodology is employed. The knowledge discovery process comprises of nine iterative and interactive steps that needs to be followed when applied on any of the problems. Each step is explained below as per our study objectives and illustrated with a figure as well.

3.1 Data Selection

3.2 Context and Data Acquisition

Data selection is the initial step of the process. In this step, data is selected and segregated into meaningful datasets as per the research objectives. The purpose of this research is to predict whether a household of a country is poor or not. The data for this study has been acquired from the world bank data repository (). The dataset for this study includes data and metadata from surveys for three developing countries and therefore, for each of the country we have data at household and individual level. The dataset contains six files each for individual and household, training and testing data of the countries.

3.3 Data Description

The dataset contains survey data for three individual countries A, B and C. Each household in a country is identified by a unique *id* and each individual is identified by the variable *iid*. Each column of the dataset, are actually the survey questions encoded into variables and the responses for the variables of each household are actually the rows. Each question is either a multiple-choice question, where the choices are stored as individual strings or numerical value. There are six training files in total (figure 5)

The dataset has been structured in such a way that the values of id in household cross match with individual datasets. Here's a glimpse of how the household and individual training data looks like:

	wBXbHZmp	SlDKnCuu	AlDbXTlZ	...	poor
id					
80389	JhtDR	GUusz	aQelm	...	True
9370	JhtDR	GUusz	ceclq	...	True
39883	JhtDR	GUusz	aQelm	...	False
18327	JhtDR	alLXR	ceclq	...	True
88416	JhtDR	GUusz	ceclq	...	True

Figure 3: Training Data of Household

		HeUgMnzF	CaukPfUC	xqUooaNJ	...	poor
id	iid					
80389	1	XJsPz	mOlYV	dSJoN	...	True
	2	XJsPz	mOlYV	JTCKs	...	True
	3	TRFeI	mOlYV	JTCKs	...	True
	4	XJsPz	yAyAe	JTCKs	...	True
9370	1	XJsPz	mOlYV	JTCKs	...	True

Figure 4: Training Data of Individual

These six files contain information about the survey data of each country separately. For both the datasets, we have a binary variable 'poor', the target variable of the study that indicates whether the household and individual is below the poverty line or not. These files are listed as below:

1. *A_hhold_train.csv*: This file contains the household training data for country 'A'.
2. *B_hhold_train.csv*: This file contains the household training data for country 'B'.
3. *C_hhold_train.csv*: This file contains the household training data for country 'C'.
4. *A_indiv_train.csv*: This file contains the individual training data for country 'A'.
5. *B_indiv_train.csv*: This file contains the individual training data for country 'B'.
6. *C_indiv_train.csv*: This file contains the individual training data for country 'C'.

Country	Filename	Survey Level
A	A_hhold_train.csv	household
B	B_hhold_train.csv	household
C	C_hhold_train.csv	household
A	A_indiv_train.csv	individual
B	B_indiv_train.csv	individual
C	C_indiv_train.csv	individual

Figure 5: Structure of Data

3.4 Data Pre-Processing

3.4.1 Importing the Data

The data will be first uploaded to the google drive to use in the colab notebook. The .csv files would be loaded into the separate panda's data frames for pre-processing and cleaning.

3.4.2 Exploratory Data Analysis

The data needs to be prepared for binary classification such that it is in a format suitable for the machine learning algorithms that we have chosen to apply. The survey data that we are going to use is mostly cleaned, but there are still a few issues that we will deal with. The NULL values in the data are removed and the columns are renamed as per the requirement.

Exploratory data analysis is done to see if the data is balanced and have a brief look at how the variables are related with each other. The relationship between the features and the target variables can be inspected by plotting a distribution graph of features against the target. The distribution has been done for numerical as well as multiple choice features of the dataset. (Ref to fig 6 and 7)

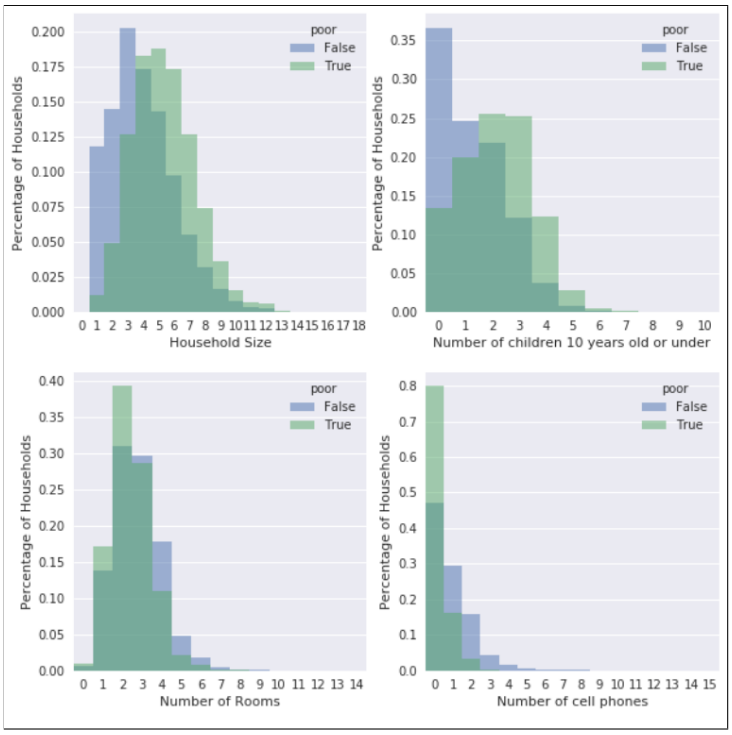

Figure 6: Distribution of the binary target variables for different numerical survey questions

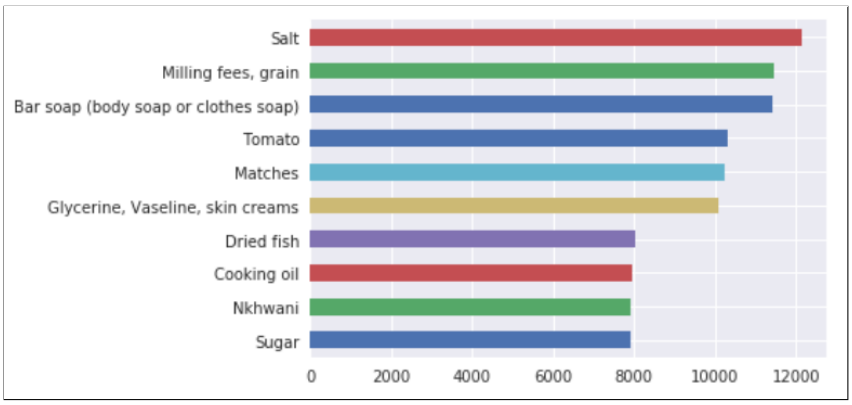

Figure 7: Distribution of the binary target variables for different multiple choice survey questions

3.5 Data Transformation

3.5.1 Feature Engineering

This is the key step of the research. In order to consider the interpreting features, feature engineering would be one of the most important and challenging part in Data Transformation. We have been provided with both household and individual level data, even though only household data should be considered for predictions, may be because we can construct additional features from the individual level data. Therefore, during feature engineering, the unnecessary features would be removed, modified while adding new features from the individual data. Specifically, we fitted a model to the original group of features and also to the combination of features we wanted to test. After that, evaluations were performed to the random permutations of features to observe their effects on the models' predictions.

3.6 Model Implementation and evaluation

The goal of the study is to predict whether a household is poor or not, hence it is a binary classification problem. As mentioned in section 2.4 we would be applying an ensemble of models using Gradient Boosting and Neural Networks. Different combinations of the mentioned machine learning algorithms would be ensembled and applied for each country. The performance of the classification model would be checked using mean log loss. It would measure the divergence of the predicted probability with the actual values for each country. The mean of the log loss scores would be the overall score of a country, and the less would be the score the more predictable is the model.

4 Proposed Design Architecture

A three-tier architecture is a type of software architecture that contains three levels/layers of computing. This study contains three phases of the proposed design architecture. The methodology explained in the previous section is implemented using these layers. All the three layers are described as below:

4.1 Database Layer

Database Layer is the first and the most essential step of the 3-tier architecture. In this layer, we handle the tasks associated with data accessing, data loading and data acquisition. For our research, the dataset will be first uploaded to the google drive for using in the colab notebook. The .csv files will be loaded into the separate panda's data frames for pre-processing and cleaning. Data is next processed and explored using different python libraries like NumPy, pandas and matplotlib. After Pre- processing and cleaning, data is transformed as per the modeling requirement. Feature engineering is used for data transformation using python packages like RDKit, pandas etc.:

4.2 Application Layer

Application layer is the second most layer of the architecture implementation. After the data is imported, cleaned and explored, next comes the model implementation i.e.; modeling. For this study, multiple combinations of ensembling Gradient boosting and

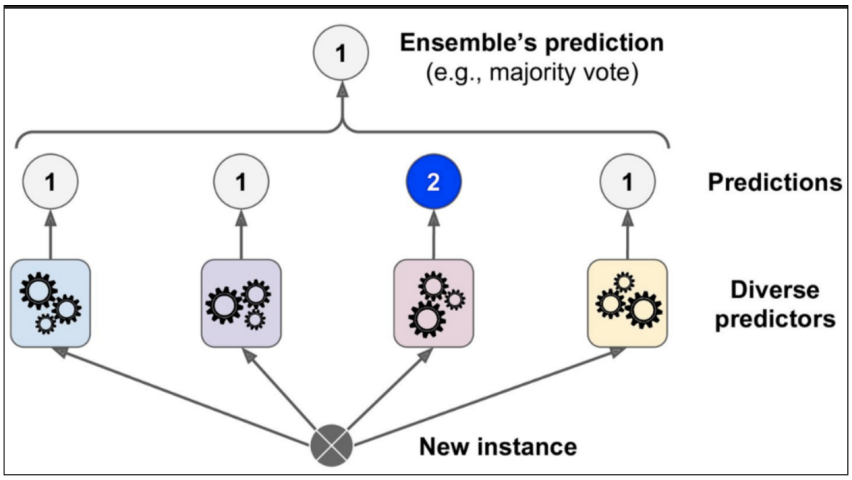

Figure 8: Working of ensemble methods

Neural networks have been applied for each country. The ensembled model has been explained briefly in figure 8. Once the ensembled model is built, training and testing of the data is done to predict the target variable, and the results are pushed forward to the next layer.

4.3 Presentation Layer

This the last but an important layer for the implementation of any methodology because it visualises the results obtained from the previous layers i.e.; Application layer. In this layer, we evaluate the performance of the model after it has been trained and tested, according to the mean log loss. Log loss scores for each country are the overall accuracy scores.

5 Proposed Implementation

This is the most crucial step of our study. In this section, we explain the architecture of the model to be implemented in the study. As mentioned in the previous section, we are going to ensemble the classifiers with the neural networks for this research as that clearly outperforms the neural network performance. Ensembled models improve the efficiency and the accuracy of the simple and complex models (Fernández-Delgado et al. (2014)). Boosting algorithms when combined with Neural networks, produce more accurate predictions as compared to single neural networks. Therefore, in this research, we implement an ensemble of models built using LightGBM Gradient Boosting and Neural Networks. Different combinations of the mentioned machine learning algorithms would be ensembled and applied for each country. For each country, we have built different combinations of models, explained as below:

1. *Country A*: 5NN + 4GB
2. *Country B*: 3NN + 4GB
3. *Country C*: 1NN + 2GB

There are many ways of combining the models for ensembling like bagging, boosting and stacking. In this study, we have bagged some models in order to reduce the variance of neural network, sampled 8 times, for 95% of the training dataset and 5% of the testing dataset. The predictions applied across the models are averaged. The proposed implementation is shown in the figure 9.

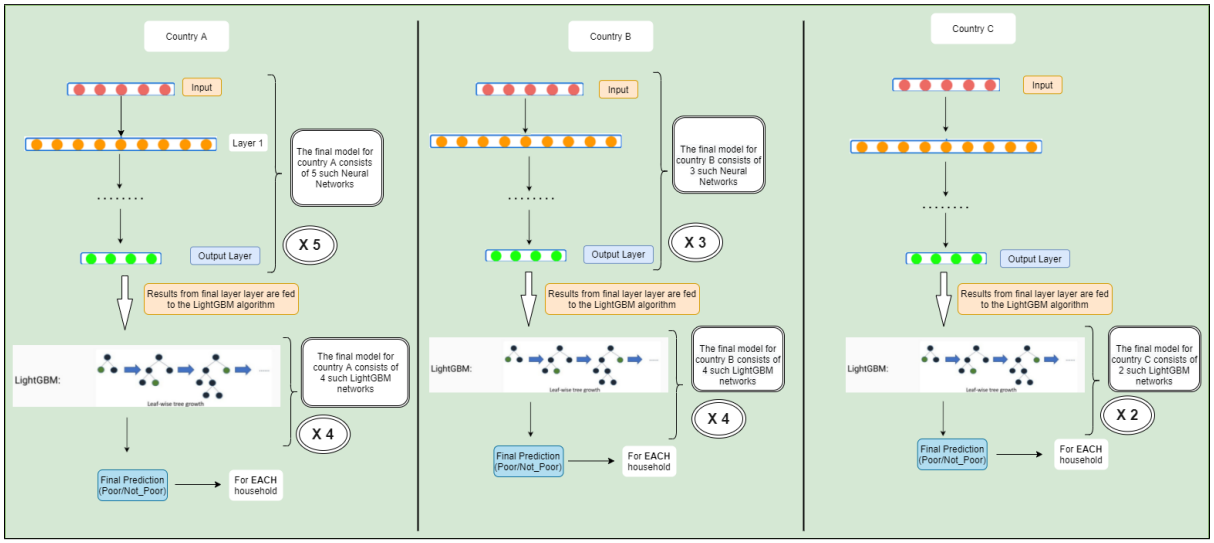

Figure 9: Proposed model architecture

6 Proposed Evaluation

For the implementation of the research, we have decided to use a different combination of ensembled models (Gradient Boosting and Neural Networks) for each country, depending on the complexity of the data. K-fold cross-validation is a resampling procedure, primarily used for the features or parameters approximation in a model. In this study, we would be implying K-fold cross-validation for considering the weight of each model in the ensemble. For each country, translation of the parameters using K-fold would turn them into an optimal and better fit. After the models are implemented, the obtained results need to be evaluated. The performance for the prediction model would be measured by the mean log loss. Log loss is the most important metric for classification algorithms. The lower is the value of log loss, the better is the predictions. For this study, we would generate log loss for each country and calculate the mean of the resultant scores, which would become the overall score of the prediction.

7 Proposed Project Management

7.1 Introduction

Project management is one of the most important aspect of any research. The foremost advantage of project management in this study is that, since the research is spread over many activities it is important to have a management of all of activities with time and resources. Gantt chart is a horizontal bar graph chart used to represent a project plan over time, in the process of project management. It is extremely useful as it helps us

to simplify the complex project into an easy follow up plan and keep a track of all the deadlines of the research. The Gantt chart for the above study is proposed in section ().

7.2 Gantt Chart

The Gantt chart for the above research is explained as below:

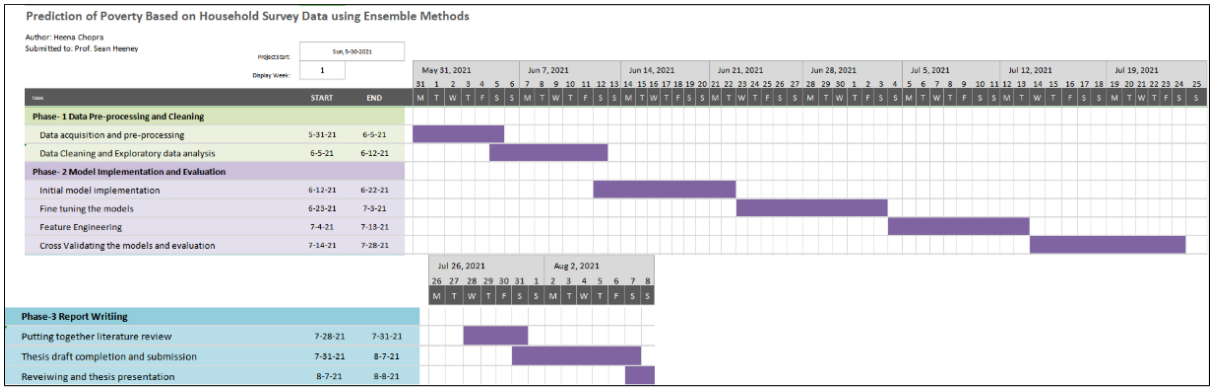

Figure 10: Proposed Gantt Chart

8 Conclusion

Data limitations are a big complication for poverty estimation and measurement. In this study, poverty is measured for developing countries at household and individual level. The main aim of the study is to predict whether a household of a country is poor or not. Feature selection,one of the most important aspect of data transformation is employed in our research to filter the features which account for the most variance in the target variable. The features selected are then quantified using the k-fold cross validation procedure. The implementation of the research is done using ensemble of neural networks with a classifier model i.e; LightGBM. The performance of the model is measured using Mean Log loss. Lower is the log loss score, better is the prediction.

References

Alsharkawi, A., Al-Fetyani, M., Dawas, M., Saadeh, H. and Alyaman, M. (2021). Poverty classification using machine learning: The case of jordan, *Sustainability* **13**(3): 1412.

Ayush, K., Uzkent, B., Burke, M., Lobell, D. and Ermon, S. (2020). Efficient poverty mapping using deep reinforcement learning, *arXiv preprint arXiv:2006.04224* .

Babenko, B., Hersh, J., Newhouse, D., Ramakrishnan, A. and Swartz, T. (2017). Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in mexico, *arXiv preprint arXiv:1711.06323* .

Blumenstock, J., Cadamuro, G. and On, R. (2015). Predicting poverty and wealth from mobile phone metadata, *Science* **350**(6264): 1073–1076.

Blumenstock, J. E. (2016). Fighting poverty with data, *Science* **353**(6301): 753–754.

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

Caldés, N., Coady, D. and Maluccio, J. A. (2006). The cost of poverty alleviation transfer programs: a comparative analysis of three programs in latin america, *World development* **34**(5): 818–837.

Choi, H. and Varian, H. (2012). Predicting the present with google trends, *Economic record* **88**: 2–9.

Eagle, N., Macy, M. and Claxton, R. (2010). Network diversity and economic development, *Science* **328**(5981): 1029–1031.

Elbers, C., Lanjouw, J. O. and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality, *Econometrica* **71**(1): 355–364.

Engelmann, G., Smith, G. and Goulding, J. (2018). The unbanked and poverty: Predicting area-level socio-economic vulnerability from m-money transactions, *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 1357–1366.

Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?, *The journal of machine learning research* **15**(1): 3133–3181.

Ferreira, M. B., Pinto, D. C., Herter, M. M., Soro, J., Vanneschi, L., Castelli, M. and Peres, F. (2020). Using artificial intelligence to overcome over-indebtedness and fight poverty, *Journal of Business Research* .

Fox, L. and Sohnesen, T. P. (2016). Household enterprises and poverty reduction in sub-saharan africa, *Development Policy Review* **34**(2): 197–221.

Grosh, M. and Baker, J. L. (1995). Proxy means tests for targeting social programs, *Living standards measurement study working paper* **118**: 1–49.

Henderson, J. V., Storeygard, A. and Weil, D. N. (2012). Measuring economic growth from outer space, *American economic review* **102**(2): 994–1028.

Hernandez, M., Hong, L., Frias-Martinez, V. and Frias-Martinez, E. (2017). *Estimating poverty using cell phone data: evidence from Guatemala*, The World Bank.

Hersh, J., Engstrom, R. and Mann, M. (2020). Open data for algorithms: mapping poverty in belize using open satellite derived features and machine learning, *Information Technology for Development* pp. 1–30.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty, *Science* **353**(6301): 790–794.

Kondmann, L. and Zhu, X. X. (2020). Measuring changes in poverty with deep learning and satellite imagery.

Kshirsagar, V., Wieczorek, J., Ramanathan, S. and Wells, R. (2017). Household poverty classification in data-scarce environments: a machine learning approach, *arXiv preprint arXiv:1711.06813* .

Li, G., Cai, Z., Liu, X., Liu, J. and Su, S. (2019). A comparison of machine learning approaches for identifying high-poverty counties: Robust features of dmsp/ols nighttime light imagery, *International journal of remote sensing* **40**(15): 5716–5736.

Masoomali, F., Isabelle, T., Ardie, O., Sy, S., Vedran, S., Garcia-Herranz, M. and Weber, I. (2020). Mapping socioeconomic indicators using social media advertising data, *EPJ Data Science* **9**(1).

McBride, L. and Nichols, A. (2015). Improved poverty targeting through machine learning: An application to the usaid poverty assessment tools, *Unpublished manuscript. Available at: http://www. econthatmatters. com/wp-content/uploads/2015/01/improvedtargeting_21jan2015. pdf* .

Michalopoulos, S. and Papaioannou, E. (2013). Pre-colonial ethnic institutions and contemporary african development, *Econometrica* **81**(1): 113–152.

Mohamud, J. H. and Gerek, O. N. (2019). Poverty level characterization via feature selection and machine learning, *2019 27th Signal Processing and Communications Applications Conference (SIU)*, IEEE, pp. 1–4.

NASA (2021). Landsat 7. [Online; accessed April 4, 2021].
URL: *https://landsat.gsfc.nasa.gov/landsat-7*

Ni, Y., Li, X., Ye, Y., Li, Y., Li, C. and Chu, D. (2020). An investigation on deep learning approaches to combining nighttime and daytime satellite imagery for poverty prediction, *IEEE Geoscience and Remote Sensing Letters* .

Nischal, K., Radhakrishnan, R., Mehta, S. and Chandani, S. (2015). Correlating nighttime satellite images with poverty and other census data of india and estimating future trends, *Proceedings of the Second ACM IKDD Conference on Data Sciences*, pp. 75–79.

Niu, T., Chen, Y. and Yuan, Y. (2020). Measuring urban poverty using multi-source data and a random forest algorithm: A case study in guangzhou, *Sustainable Cities and Society* **54**: 102014.

Omrani, H., Gerber, P. and Bousch, P. (2009). Model-based small area estimation with application to unemployment estimates, *World Academy of Science, Engineering and Technology* **49**: 793–800.

Ozgur, A. (2004). Supervised and unsupervised machine learning techniques for text document categorization, *Unpublished Master's Thesis, İstanbul: Boğaziçi University* .

Pandey, S., Agarwal, T. and Krishnan, N. C. (2018). Multi-task deep learning for predicting poverty from satellite images, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

Perez, A., Ganguli, S., Ermon, S., Azzari, G., Burke, M. and Lobell, D. (2019). Semi-supervised multitask learning on multispectral satellite images using wasserstein generative adversarial networks (gans) for predicting poverty, *arXiv preprint arXiv:1902.11110* .

Piaggesi, S., Gauvin, L., Tizzoni, M., Cattuto, C., Adler, N., Verhulst, S., Young, A., Price, R., Ferres, L. and Panisson, A. (2019). Predicting city poverty using satellite imagery, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 90–96.

Pinkovskiy, M. and Sala-i Martin, X. (2016). Newer need not be better: evaluating the penn world tables and the world development indicators using nighttime lights, *Technical report*, National Bureau of Economic Research.

Pokhriyal, N. and Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping, *Proceedings of the National Academy of Sciences* **114**(46): E9783–E9792.

Poverty (2021). [Online; accessed April 4, 2021].
URL: *https://en.wikipedia.org/wiki/Povertycite$_n$ote − OWiD$_E$P −*

Rasmussen, C. E. (2003). Gaussian processes in machine learning, *Summer school on machine learning*, Springer, pp. 63–71.

Sani, N. S., Rahman, M. A., Bakar, A. A., Sahran, S. and Sarim, H. M. (2018). Machine learning approach for bottom 40 percent households (b40) poverty classification, *Int. J. Adv. Sci. Eng. Inf. Technol* **8**(4-2): 1698.

Schreiner, M. (2010). A simple poverty scorecard for pakistan, *Journal of Asian and African Studies* **45**(3): 326–349.

Subash, S., Kumar, R. R. and Aditya, K. (2018). Satellite data and machine learning tools for predicting poverty in rural india, *Agricultural Economics Research Review* **31**(347-2019-571): 231–240.

Tingzon, I., Orden, A., Go, K., Sy, S., Sekara, V., Weber, I., Fatehkia, M., García-Herranz, M. and Kim, D. (2019). Mapping poverty in the philippines using machine learning, satellite imagery, and crowd-sourced geospatial information, *AI for Social Good ICML 2019 Workshop*.

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting elections with non-representative polls, *International Journal of Forecasting* **31**(3): 980–991.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S. and Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in africa, *Nature communications* **11**(1): 1–11.

Yin, J., Qiu, Y. and Zhang, B. (2021). Identification of poverty areas by remote sensing and machine learning: A case study in guizhou, southwest china, *ISPRS International Journal of Geo-Information* **10**(1): 11.