Thailand and The World Economy

P-ISSN: 2630-0931
E-ISSN: 2651-0529

Homepage : *https://www.tci-thaijo.org/index.php/TER*

Vol. 43, No. 2, May – August 2025 Page [138–162]

Selection of the Best Machine Learning Model to Predict Poverty Conditions: A Study on North-Eastern Wetland Region of Bangladesh

Md. Rashidul Hasan[*]
Ph.D. Research Scholar, Department of Statistics,
Shahjalal University of Science & Technology, Bangladesh.
Md. Zakir Hossain
Professor, Department of Statistics,
Shahjalal University of Science & Technology, Bangladesh.

Received 3 October 2023, Received in revised form 28 April 2024,
Accepted 13 August 2024, Available online 7 May 2025

Abstract

Machine learning (ML) algorithms are effective techniques for predicting households' poverty conditions so that they might benefit from poverty alleviation programs. The study's primary objective is to find out the determinants of poverty and select the best ML model to predict the poverty conditions of the north-eastern wetland region of Bangladesh. This study used data from 2340 households that were collected through a household survey by a research project sponsored by the GARE Program, Ministry of Education, GoB. The multiple logistic regression (MLR) model was employed to extract the factors associated with household poverty. Six ML algorithms, including support vector machine, Naïve Bayes, logistic regression, K-nearest neighbor, decision tree, and random forest were applied to predict poverty conditions, and their performances were measured by using accuracy, precision, recall, F1-score, and AUROC. The study's findings show that **district, micro-credit status, household size, age,** NGO membership, marital status, **per capita income, cultivable land, electricity connection,** and **livestock ownership** are the significant determinants of wetland people's poverty. The findings also show that the support vector machine is the best model for predicting poverty level LPL with an accuracy of 82%, F1-score of 59%, and AUROC of 72%, and the logistic regression is the best model for predicting poverty level UPL with an accuracy of 81%, F1-score of 84%, and AUROC of 80%. The proposed algorithms may help improve poverty conditions by accurately predicting target poor groups. The determinants may be effective in developing policies to lessen poverty in the wetland region of Bangladesh.

Keywords: Poverty, Determinants, Prediction, Machine Learning, Bangladesh
JEL Classifications: I32, I38

[*] **Corresponding author**: Email: rashidulhasan067@gmail.com

1. Introduction

A wetland is a region or area where the soil is perpetually or sporadically wet (Ministry of Law, 2013). Floodplains, low marshes, submerged regions, riverine mudflats, open water bodies, *haors* (seasonal water bodies), *baors* (oxbow lakes), *beels* (perennial water bodies), etc. are among the several forms of wetlands in Bangladesh. *Haors* are large, seemingly bowl-shaped geological depressions that collect surface runoff water. The *haor* districts of Bangladesh occupy 19,998 square kilometers of land, or 13.56% of the country's total area. About 43% (8585 sq. km.) of the *haor* district's total area is made up of wetlands, comprising 373 *haors* (Centre for Environmental and Geographic Information Services, 2012). The basic means of subsistence are largely insufficient in the *haor* regions, notwithstanding their expertise in growing *boro* rice and freshwater fishing. According to Khondker & Mahzab (2015), people from *haor* regions are considered as belonging to the "backward section" of Bangladeshi society because they are significantly less developed than the nation's general population in regards to per capita income, consumption, electricity facilities, roads, and poverty. The lengthy seasonality of the wet monsoon contributes to the *haor* people's frequent unemployment (Hasan & Hossain, 2024). Only around 30% of *haor* people are reported to be above the upper poverty line, and roughly one-third of them are said to lie below the lower poverty line (Chowdhury, 2014). As a result, a significant portion of the *haor* people struggle with hunger and other basic needs (Kazal et al., 2017). Through a number of poverty alleviation programs, the Bangladeshi government has been trying to end poverty among its population and to meet the Sustainable Development Goals (SDGs) target. Several Non-government Organizations (NGOs) are additionally providing incentives to the poor in an effort to raise their income and improve their miserable situations (Hashemi et al., 1996).

In these circumstances, a system should be developed to determine whether poor households in wetland areas are eligible to receive poverty alleviation initiatives. This challenge can be tackled by applying machine learning techniques to predict which households fall below and above the poverty lines. In the field of machine learning, determinant identification, or feature selection, is an essential pre-processing step. Determinant identification is also essential for taking effective strategies for poverty alleviation in a region. There is an extensive corpus of literature in almost every country that employed several methods to explore the key components of poverty (for example, Acharya et al., 2022; Achia et al., 2010; Biyase & Zwane, 2017; Korankye, 2014; Ogwumike & Akinnibosun, 2013; Rhoumah, 2016; Spaho, 2014). Acharya et al. (2022) used a binary logistic regression model to determine the factors affecting poverty in Nepal. The study found several factors that were associated with the risk of poverty, including the household head's illiteracy status, remittance status, landholding status, access to the nearest market, number of literate persons of working age, etc. Using a logistic regression model, Achia et al. (2010) carried out research to pinpoint the key factors contributing to poverty in Kenya. The study used demographic and health survey (DHS) data and identified that age, educational level of household head, size of household, type of residence, religion, and ethnicity are the significant risk factors for poverty. Biyase & Zwane (2017) looked into the factors influencing household well-being and poverty in South Africa using fixed-effect and random-effect probit models. Their research revealed that household well-being, as well as poverty, were highly influenced by factors like education, sex, race, employment, and the marital status of the household head. In Rhoumah's (2016) study in Malaysia, it was shown that the three main factors influencing poverty among fishermen's households were income, education, and marital status. According to a study by Korankye (2014), the main determinants of poverty in Ghana include the prevalence of diseases, lack of education, corruption, and inefficient government. According to Spaho (2014), the results of two regression models showed that the number of household members, place of residence, and job status were the most important determinants of household poverty in Albania. A study was conducted by Ogwumike & Akinnibosun (2013) to determine what factors lead to poverty in farming households in Nigeria. As per the results of the study, the key factors that determine poverty in farming households are age, household size, income, and the quantity of farms.

Machine learning algorithms have been very popular recently as an accurate way to predict poverty levels, and their application is growing day by day (Kambuya, 2020; Li et al., 2022; Mohamud et al., 2019; Santa et al., 2023; Sohnesen et al., 2017). Nevertheless, no research has been conducted in Bangladesh to predict the poverty level of people living in wetlands using machine learning techniques. This study marks the first time that machine learning algorithms have been adopted to predict poor households using cross-sectional data. So, the study's hypothesis is to propose the best ML model for the prediction of poverty conditions in the north-eastern wetland region of Bangladesh by using an MLR features selection approach in conjunction with efficient ML algorithms. The following are some contributions made by this study:

(i) Identification of determinants: this study identified the determinants of poverty based on the p-value (<0.05) and odds ratio (OR) of the MLR model.

(ii) Machine learning system: this study predicted poverty level by applying several ML algorithms such as support vector machine, Naïve Bayes, logistic regression, KNN, decision tree, and random forest.

(iii) Performance evaluation: this study evaluated the performance of ML models based on accuracy, precision, recall, F1-score, and AUROC.

(iv) Scientific validation: this study performed stratified K-fold cross-validation upon the same dataset and compared the outcome to confirm validity.

(v) Proposed model: this study proposed the best model based on the comparison of performance metrics.

2. Literature Review

The literature has several studies (Alsharkawi, 2021; Kim, 2021; Min et al., 2022; Sani et al., 2018; Shen, 2021; Sheng, 2021; Talingdan, 2019; Wang et al., 2020; Wong, 2022; Zixi, 2021) that have attempted to predict the condition of poverty utilizing machine learning algorithms. For instance, Sani et al. (2018) used the Naïve Bayes, decision tree, and K-nearest neighbor algorithms to categorize the poverty status of the lowest 40% of households in Malaysia. A dataset from the Society Wellbeing Department's national poverty data bank was used in that study. The study also utilized a 10-fold cross-validation approach and showed that the decision tree algorithm performs better overall than the other algorithms, with an accuracy of 99.3%. A study was carried out in the Philippines by Talingdan (2019) to examine the effectiveness of household-level poverty classification algorithms. The study used five machine learning algorithms such as ID3, Naïve Bayes, decision tree, logistic regression, and K-nearest neighbor. The study found that the Naïve Bayes algorithm is an effective technique for classifying households that are poor and non-poor. Using a dataset from the Inter-American Development Bank, Wang et al. (2020) performed a study to predict the level of poverty in Costa Rica. To estimate the category placement in the dependent variable, the study employed a multinomial logistic regression model. The study also used K-means clustering, decision trees, and the gradient boosting machine (GBM) to predict the poverty level. The study's findings demonstrated that GBM offers superior prediction with an accuracy of 92.6%. Shen (2021) carried out a study based on the Costa Rican poverty dataset and applied logistic regression, support vector machine, K-nearest neighbor, decision tree, and random forest algorithms to classify poor households. The study's findings showed that the decision tree algorithm performs well, with an average accuracy of 89.0%.

To predict the level of poverty, Zixi (2021) used machine learning algorithms on multidimensional poverty index data from several countries. Lasso regression was employed in that study to identify the covariates of poverty. The study also applied four machine learning algorithms, namely decision tree, random forest, gradient boosting, and artificial neural network, and found that gradient boosting is the best algorithm for predicting poverty with an accuracy of 78.5%. Data from the Household Expenditure and Income Survey (HEIS) was used by Alsharkawi (2021) to determine and quantify the poverty condition of Jordanian households. Several machine learning classification models, such as logistic regression, ridge regression, stochastic gradient descent, passive aggressive, K-nearest neighbor, decision tree, extra tree, support vector machine, Naïve Bayes (NB), Ada boost, bagged decision trees, random forest, GBM, light GBM, and scalable tree boosting system, were used in that study. The study's findings revealed that, with an F1-score of 81.0%, the light GBM algorithm performed best. In order to assist the business and government sectors, Kim (2021) conducted a study that uses two supervised machine learning algorithms, namely random forest and gradient boosted trees, to predict Costa Rican households' poverty level. The algorithms used in the study produced accuracy rates of 78.1% and 79.6%, respectively.

Sheng (2021) conducted research at Chuzhou University in China on 5,000 underprivileged college students. The study employed principal component analysis (PCA) to extract the features and various classification algorithms such as K-nearest neighbor, support vector machine, Gaussian NB, logistic regression, linear discriminant analysis, classification and regression tree, extreme gradient boosting, and random forest (RF) to confirm the superiority of the RF-PCA dimensionality reduction. The study found that random forest performs better than other models, with an accuracy rate of 78.6%. A study by Min et al. (2022) used linear regression, decision trees, and random forest algorithms to predict the level of poverty, where the algorithms were assessed using the poverty dataset of Costa Rica. The study also applied the Boruta feature selection approach when making predictions. The experimental findings of the study concluded that random forest performs best with R^2 and RMSE scores of 0.946 and 0.259, respectively. Wong (2022) conducted a study to address the global issue of poverty using the Demographic and Health Survey (DHS) 2014 data of Cambodia for the machine learning model. The study applied softmax, random forest, and artificial neural network (ANN) classifiers and compared them. The study found that ANN, with an accuracy of 87.0%, produces better results when compared with other models.

3. Materials and Methods

3.1 Study area

The study was conducted in the north-eastern wetland region of Bangladesh, covering six *haor*-prone districts such as Sunamganj, Sylhet, Habiganj, Maulvibazar, Netrokona, and Kishoreganj. *Haors* are mainly found in the districts of Sunamganj, Sylhet, Netrokona, and Kishoreganj. There are 366 *haors* in the aforementioned six districts, although only seven are located in the Brahmanbaria district (Centre for Environmental and Geographic Information Services, 2012).

3.2 The data

The required data for the study was taken from the data collected through a household survey (conducted during February-December 2019) by a research project funded by the Grants for Advanced Research in Education (GARE) Program, Ministry of Education, Government of Bangladesh (GoB). The data is cross-sectional because it was collected from several individuals at a single time point.

3.3 Sample design

A cluster-sampling design was used in the survey from which the data was extracted, and *haor* attached unions were considered as clusters. The survey covers a total of 30 clusters. The sample size for the survey was 2340, according to the standard sample size determination formula[1]. The survey used the following procedures to select clusters and households:

(i) The number of *haors* in each of the six districts is defined and determined.

(ii) A stratified random sampling with proportional allocation was employed to estimate the number of *haors* in each district (stratum). A systematic probability proportional to size (PPS) sampling was then employed to select *haor* from each of the districts.

(iii) From each of the chosen *haor*, a cluster was chosen randomly.

(iv) The households within the cluster were chosen at random using the UNICEF pencil-spin method.

(v) Finally, a total of 2340 households (78 from each cluster) were chosen from 30 clusters for interview.

3.4 Estimation of poverty lines

Generally, two methods are applied to estimate the household-level poverty lines. The first one is the Cost of Basic Needs (CBN) method and the other one is the Direct Calorie Intake (DCI) method. In this study, the CBN method is utilized to calculate the poverty lines. The CBN method is recommended by the World Bank and used by planners, policymakers, and international agencies (Bangladesh Bureau of Statistics, 2017). The Bangladesh Bureau of Statistics has been applying the CBN method to estimate the incidence of poverty since 1995–96 (Bangladesh Bureau of Statistics, 2023). This method estimates two poverty lines, (i) the lower and (ii) the upper poverty line, in three steps.

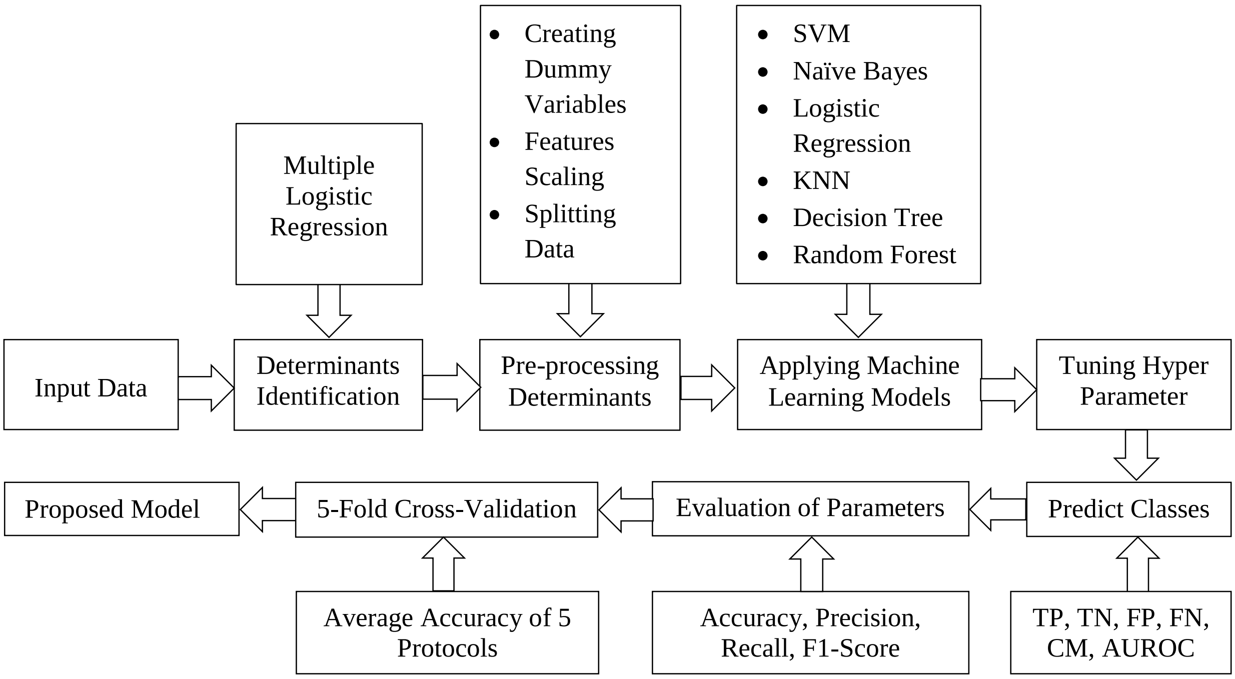
The first step involved the calculation of the food poverty line. The second step involved the calculation of two non-food allowances for non-food consumption. In the third step the lower poverty line is calculated by adding non-food lower allowance with the food poverty line and the upper poverty line is calculated by adding non-food upper allowance with the food poverty line.

According to the CBN method, a household is considered to be below the lower poverty line (LPL) if its annual per capita consumption expenditure is less than Taka 16296.5 and below the upper poverty line (UPL) if it is less than Taka 21638.2.

[1] $n = \frac{p(1-p)Z^2}{(0.04p)^2} \times Deff$; Where, p = percentage indicator, Z = normal variate value with 95% confidence interval, $0.04p$ = relative error margin, and $Deff$ = design effect.

3.5 The overall machine-learning system

Figure 1: Diagrammatic Representation of Machine Learning System



Note: TP=True Positive, TN=True Negative, FP=False Positive, FN=False Negative
Source: Author's summarization.

Figure 1 shows the machine learning system's overall diagrammatic representation. First, we inputted data and eliminated unnecessary observations from the analysis. Secondly, the multiple logistic regression (MLR) model was employed to extract the determinants. Thirdly, pre-processed determinants by (i) creating a dummy variable for each value in each categorical variable, (ii) transforming all the numeric features on a comparable scale using a standardization technique, and (iii) splitting the dataset into training and test sets. The training set consists of 80% (*N*=1872) of the data, while the remaining 20% (*N*=468) is reserved for testing. Fourthly, used several supervised type ML classifiers and tuned their hyper parameters. Fifthly, predicted poverty levels based on the confusion matrix (CM) and area under the ROC (AUROC) curve. Sixthly, the accuracy, precision, recall, and F1-score were calculated to evaluate the performance of classifiers. Finally, the study applied stratified 5-fold cross-validation and identified the best model.

3.5.1) Determinants/Features selection technique

This study developed two multiple logistic regression (MLR) models to identify the determinants associated with household poverty conditions LPL and UPL. The models are stated below:

Let, $X = (X_1, X_2, ..., X_{12})'$ is a vector of the collection of predictors and $Y_i(LPL)$ is a binary outcome variable that indicates the household's poverty condition based on LPL.

Where, $\mathrm{Y_i}(LPL) = \begin{cases} 1 & \text{if the i - th household lies below the LPL} \\ 0 & \text{otherwise} \end{cases}$

The conditional probability of the i-th household lies below LPL given X be written as

$$\pi_i(LPL) = \text{Prob}\left[\text{Y}_\text{i}(LPL) = 1 \mid X\right] = \frac{\exp\left(\beta_0 + \sum_{i=1}^{12} \beta_i X_i\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{12} \beta_i X_i\right)}$$

Here, $\beta = (\beta_1, \beta_2, \ldots\ldots, \beta_{12})$ is a vector of unknown parameters ordinarily estimated by the method of maximum likelihood.

The logit of $\pi_i(LPL)$ with predictors is given by

$$\log_e\left(\frac{\pi_i(LPL)}{1-\pi_i(LPL)}\right) = \beta_0 + \sum_{i=1}^{12} \beta_i X_i \qquad (1)$$

Similarly, the logit of $\pi_i(UPL)$ with predictors is given by

$$\log_e\left(\frac{\pi_i(UPL)}{1-\pi_i(UPL)}\right) = \beta_0 + \sum_{i=1}^{12} \beta_i X_i \qquad (2)$$

Where, $\pi_i(UPL)$ is the conditional probability of the i-th household lies below UPL given X. Models (1) and (2) are known as multiple logistic regression models. This study considered several individual-level and household-level characteristics as explanatory variables in models based on previous research (Borko, 2017; Imam et al., 2018; Kazal et al., 2017) which are presented in Table 1.

Table 1: Explanatory Variables with their Symbol, Description, Types, and Class Level

Variables	Symbol	Description	Types	Class Level
District	X_1	District of households	Categorical	Sunamganj, Sylhet, Habiganj, Netrokona, Kishoreganj
Micro-Credit Status	X_2	Households' micro-credit status	Categorical	Non-borrower, Borrower
Household Size	X_3	Number of household members	Categorical	<4, 4 or more
Age	X_4	Age (in year) of household heads	Numerical	Measured in number
Gender	X_5	The gender of household heads	Categorical	Male, Female
Occupation	X_6	Occupation of household heads	Categorical	Farming, Day laborer, Off-farm activities, Service/Business, Household work, Others
Marital Status	X_7	Marital status of household heads	Categorical	Married, Unmarried, Widowed/Divorced
NGO Membership	X_8	Households' NGO membership	Categorical	No, Yes
Per Capita Income	X_9	Households' per capita income (in Taka)	Categorical	≤10000, 10000-20000, 20000-30000, 30000 or more
Cultivable Land	X_{10}	Households' cultivable land (in decimal)	Categorical	No land, 1-15, 16-50, 50+

Variables	Symbol	Description	Types	Class Level
Electricity Connection	X_{11}	Access of electricity in households	Categorical	No, Yes
Livestock Ownership	X_{12}	Livestock ownership of households	Categorical	No, Yes

Source: Pre-existing literature (Borko, 2017; Imam et al., 2018; Kazal et al., 2017).

3.5.2) Machine learning algorithms

In our study, six supervised machine learning algorithms were used to predict the household's poverty level based on significant determinants found in the MLR models. Following are descriptions of the algorithms:

3.5.2.1 Support vector machine

The support vector machine (SVM) was first developed by Vapnik (Cortes & Vapnik, 1995). It can be utilized to predict households with poverty conditions where the output variable is categorical (Alsharkawi et al., 2021).

Let us consider a training data matrix $D_{ij} = (x_i, y_i)$; $i = 1,2,..., m = 1872$; $j = 1,2,..., n = 22$. Where, $x_i = (x_{i1}, x_{i2},..., x_{in})^T$ is an input vector of features and $y_i = (y_{i1})^T$ is the outcome variable that takes a value "1" if the household lies below the poverty line, and "0" otherwise. The SVM considers a linear function with the following hyper-plane: $f(x_i) = w^T x_i + b$

Where, w = normalized weight vector and b = bias of the linear classification. The following expression divides the data into two groups, 1 and 0 if the data is linearly separable.

$$y_i = \begin{cases} 1 & \text{if } w^T x_i + b \geq 0 \\ 0 & \text{if } w^T x_i + b < 0 \end{cases}$$

We employ the SVM kernel to easily separate the classes if the data is unable to separate linearly. In this instance, the hyper-plane is $f(x_i) = \sum_{n=1}^{N} \alpha_n y_n K(x_n, x_i) + b$

Where, α_n = Lagrange multiplier, y_n = membership class label, $K(x_n, x_i)$ = kernel function between x_n and x_i. This study used the radial basis kernel function (RBF) for SVM. The mathematical expression for RBF kernel is $k(x_i, y_i) = \exp(-\gamma \|x_i - y_i\|^2)$ for $\gamma > 0$.

3.5.2.2 Naïve Bayes algorithm

The Naïve Bayes (NB) algorithm is a straightforward probabilistic algorithm that constructs a classifier using the Bayes rule and a set of conditional independence assumptions. In 1973 (Duda et al., 1973), the Naïve Bayes algorithm was first presented. It was then reintroduced in 1992 (Langley et al., 1992).

Let $X = (X_1, X_2,..., X_n)$ is a vector of features and Y is the outcome variable that takes a value "1" if the household lies below the poverty line, and "0" otherwise. Then the classifier will calculate the posterior probability across all possible values of Y , for each new input X that we ask the classifier to classify. According to the Bayes rule, the probability that Y will take the value 1 given $X_1, X_2,..., X_n$ would be

$$P(Y=1\,|\,X_1,X_2,...,X_n)=\frac{P(Y=1)P(X_1,X_2,...,X_n\,|\,Y=1)}{P(X_1,X_2,...,X_n)} \tag{3}$$

Assuming that the X's are conditionally independent i.e., $P(X_1,X_2,...,X_n)=\prod_i P(X_i)$.

Then we can rewrite (3) as $P(Y=1\,|\,X_1,X_2,...,X_n)=\dfrac{P(Y=1)\prod_i P(X_i\,|\,Y=1)}{\prod_i P(X_i)}$

Similarly, the probability that Y will take the value 0 given $X_1,X_2,...,X_n$ would be

$$P(Y=0\,|\,X_1,X_2,...,X_n)=\frac{P(Y=0)\prod_i P(X_i\,|\,Y=0)}{\prod_i P(X_i)}$$

The Naïve Bayes classifier then puts a feature X_i into the class 1 if and only if $P(Y=1\,|\,X_i)>P(Y=0\,|\,X_i)$ otherwise puts it into the class 0. According to Talingdan (2019), the Naïve Bayes algorithm is a useful method for predicting poor and non-poor households in the Philippines.

3.5.2.3 Logistic regression classifier

The logistic regression (LR) classifier can be applied to categorize an observation into two or more categories (Myers et al., 2012), but in our study, we focused on the common binary response version.

Let us consider a training data matrix $D_{ij}=(X,Y)$; $i=1,2,...,m=1872$; $j=1,2,...,n=22$. Where, $X=(x_{i1},x_{i2},...,x_{in})^T$ is an input vector of features and $Y=(y_{i1})^T$ is the outcome variable that takes a value "1" if the household lies below the poverty line and "0" otherwise. Then, the logistic regression model can be written as, $P(Y=1\,|\,X)=\sigma(Z)$

Where, $\sigma(Z)=\dfrac{1}{1+e^{-Z}}$ and $Z=b+\sum_{i=1}^{n}W_iX_i$.

Here W_i = real-valued weight matrix and, b = bias term called the intercept.

Similarly, $P(Y=0\,|\,X)=\dfrac{1}{1+e^{Z}}$

Now the following expression divides the data into two classes: 1 and 0.

$$Y=\begin{cases}1 & \text{if } \sigma(Z)\geq 0.5\\ 0 & \text{if } \sigma(Z)<0.5\end{cases}$$

Alsharkawi et al. (2021) applied the logistic regression classifier on household expenditure and income survey data to classify poverty in Jordan.

3.5.2.4 K-nearest neighbor algorithm

The K-nearest neighbor (KNN) algorithm was first presented by Fix and Hodges in 1951 (Fix & Hodges, 1951), while Cover & Hart (1967) later developed it in 1967. The primary goal of this algorithm is to categorize new features in test data using the features of the input training data. The new features of test data were classified into the category of outcome variable by the majority of the categories' K-nearest neighbors.

Let us consider a training data matrix $D_{ij}=(X,Y)$; $i=1,2,...,m=1872$; $j=1,2,...,n=22$. Where, $X=(x_{i1},x_{i2},...,x_{in})^T$ is an input vector of features and $Y=(y_{i1})^T$

is the outcome variable that takes a value "1" if the household lies below the poverty line and "0" otherwise. Then, the KNN algorithm involved the following steps:

Step 1: Select the number K of the neighbors.

Step 2: Store the training data of features vector *X* and the training data of outcome variable *Y* in an n-dimensional space.

Step 3: Store the test data of features vector *X*. Calculate the Euclidean distance of p-features of a training data set by the formula $dist = \sqrt{\sum_{i=1}^{p}(a_k - b_k)^2}$. Where, a_k and b_k are the two data points of a feature.

Step 4: Take K-nearest neighbors using the rank of Euclidean distances. Then, the K-nearest classifier searches the K Euclidean distance for each test data.

Step 5: Classify the features of test data into the category of outcome variable based on the majority of category among its nearest neighbors.

Santoso et al. (2016) examined the accuracy of KNN and learning vector quantization (LVQ) in classifying the level of poverty. The result suggested that KNN performed better than the LVQ.

3.5.2.5 Decision tree classifier

It is a classifier based on a tree structure, where each internal node represents the dataset's features, and the leaf node represents the classification (Quinlan, 1986). Building a decision tree involved the following steps:

Step 1: Calculate the Gini index for outcome variable *Y* and each input feature *X*. Suppose a data set *D* contains samples from *C* classes, then the Gini index is

$gini(D) = 1 - \sum_{c=1}^{C} P_c^{\,2}$; Where, P_c = relative frequency of class *C*.

Step 2: Calculate the weighted sum of the Gini indices for each feature. The feature having the minimum weighted sum gives the maximum information. When a data set *D* divides on *S* into two subsets D_1 and D_2 then the weighted sum of Gini index is

$gini_S(D) = \frac{D_1}{D} gini(D_1) + \frac{D_2}{D} gini(D_2)$; Where, $gini(D_1) < gini(D)$, $gini(D_2) < gini(D)$

Step 3: Choose the feature with minimum weighted sum value.

Step 4: Repeat steps 1, 2, and 3 until a generalized tree has been created.

Zixi (2021) used the decision tree approach to predict poverty using multidimensional poverty index data from different countries.

3.5.2.6 Random forest classifier

Random forest is an effective ensemble-based classifier that was established by Breiman in 2001 (Breiman, 2001). It is a powerful predictive algorithm for classifying poverty conditions (Thoplan, 2014) under the following two phases:

Phase I: Create the random forest by combining decision trees:

Let us consider a training data matrix $D_{ij} = (X, Y)$; $i = 1,2,..., m = 1872$; $j = 1,2,..., n = 22$. Where, $X = (x_{i1}, x_{i2},..., x_{in})^T$ is an input vector of features and $Y = (y_{i1})^T$ is the outcome variable that takes a value "1" if the household lies below the poverty line and "0" otherwise. The following steps can then be used to show how the random forest algorithm works:

Step 1: Select *k* data points (households) at random from *m* with replacement to build five new datasets, which are also called bootstrapped datasets.

Step 2: Select *p* features at random from the *n* (*p*<*n*) available features from each bootstrapped dataset.

Step 3: Using each bootstrapped dataset along with selected features, construct five decision trees using binary recursive partitioning.

Step 4: Repeat steps 1 to 3 until a large number of trees (2340 trees) are produced to create a forest.

Phase II: Make predictions for each tree created in the first phase:

Step 1: For predicting Y at the new test data point $X=(x_{i1},x_{i2},...,x_{in})^T$; $i=1,2,...,m=468$, pass these data points through each tree one by one and note down the predictions, called base learners.

Step 2: Base learners (say) $h_1(X),h_2(X),...,h_l(X)$ are then combined to give an ensemble predictor $\hat{f}(X)=\arg\max\limits_{Y\in[0,1]}\sum\limits_{L=1}^{l} I[h_L(X)=Y]$. Here $\hat{f}(X)$ is the most frequently predicted class that wins the majority of votes.

3.6 Statistical analysis

The association between households' background characteristics and their poverty conditions was investigated using the chi-square test of independence. A z-test was used to determine whether the difference between households in poverty (those below the poverty line) and those who were not (otherwise) was significant for the continuous variable (age). Significant determinants were taken into consideration from the MLR models with p-value (<0.05). The performance of ML models was evaluated using several performance metrics such as accuracy, precision, recall, F1-score, AUROC, etc. SPSS version 25.0 and Google Colaboratory for Python were used for analyses.

4. Empirical Results

This section presents the empirical results of the analysis. Sub-section 4.1 describes the poverty conditions by background characteristics of the households and household heads, sub-section 4.2 identifies the determinants of household poverty conditions, and sub-section 4.3 describes the performance evaluation of several machine learning algorithms.

4.1 Poverty conditions by background characteristics of the households and household heads

The association of poverty conditions with several background characteristics of the households and household heads has been carried out to study the differentials of poverty. The significance of the variables was examined through the *p*-values and given in Table 2.

Table 2: Association of Covariates with Poverty Conditions Based on LPL and UPL

Characteristics	**Overall, N(%)**	**Poverty Condition (Based on LPL)**		**p-value**	**Poverty Condition (Based on UPL)**		**p-value**
		HHs Below LPL, n(%)	**Otherwise, n(%)**		**HHs Below UPL, n(%)**	**Otherwise, n(%)**	
Total	**2340(100)**	**578(24.7)**	**1762(75.3)**		**1290(55.1)**	**1050(44.9)**	
District							
Sunamganj	1256(53.7)	316(54.7)	940(53.3)	<0.001	714(53.3)	542(51.6)	<0.001
Sylhet	78(3.3)	4(0.7)	74(4.2)		23(1.8)	55(5.2)	
Habiganj	312(13.3)	82(14.2)	230(13.1)		201(15.6)	111(10.6)	
Netrokona	379(16.2)	69(11.9)	310(17.6)		183(14.2)	196(18.7)	
Kishoreganj	315(13.5)	107(18.5)	208(11.8)		169(13.1)	146(13.9)	
Micro-Credit Status							
Non-Borrower	733(31.3)	230(39.8)	503(28.5)	<0.001	468(36.3)	265(25.2)	<0.001
Borrower	1607(68.7)	348(60.2)	1259(71.5)		822(63.7)	785(74.8)	
Household Size							
<4	306(13.1)	13(2.2)	293(16.6)	<0.001	88(6.8)	218(20.8)	<0.001
4 or more	2034(86.9)	565(97.8)	1469(83.4)		1202(93.2)	832(79.2)	
Gender							
Male	1796(78.6)	482(83.4)	1314(74.6)	<0.001	1020(79.1)	776(73.9)	0.003
Female	544(23.2)	96(16.6)	448(25.4)		270(20.9)	274(26.1)	
Occupation							
Farming	461(19.7)	92(15.9)	369(20.9)	<0.001	233(18.1)	228(21.7)	<0.001
Day laborer	532(22.7)	168(29.1)	364(20.7)		329(25.5)	203(19.3)	
Off-farm activities	308(13.2)	116(20.1)	192(10.9)		221(17.1)	87(8.3)	
Service/ Business	441(18.8)	85(14.7)	356(20.2)		204(15.8)	237(22.6)	
Household work	442(18.9)	74(12.8)	368(20.9)		214(16.6)	228(21.7)	
Others	156(6.7)	43(7.4)	113(6.4)		89(6.9)	67(6.4)	
Marital Status							
Married	2158(92.2)	549(95.0)	1609(91.3)	0.004	1192(92.4)	966(92.0)	0.369
Unmarried	60(2.6)	5(0.9)	55(3.1)		28(2.2)	32(3.0)	
Widowed/ Divorced	122(5.2)	24(4.2)	98(5.6)		70(5.4)	52(5.0)	
NGO Membership							
No	1143(48.8)	359(62.1)	784(44.5)	<0.001	701(54.3)	442(42.1)	<0.001
Yes	1197(51.2)	219(37.9)	978(55.5)		589(45.7)	608(57.9)	
Per Capita Income (in Taka) Per Year							
≤10000	197(8.4)	148(25.6)	49(2.8)	<0.001	117(13.8)	19(1.8)	<0.001
10000-20000	1190(50.9)	413(71.5)	777(44.1)		948(73.5)	242(23.0)	
20000-30000	617(26.4)	13(2.2)	604(34.4)		138(10.7)	479(45.6)	
30000 or more	336(14.4)	4(0.7)	332(18.8)		26(2.0)	310(29.5)	

Characteristics	Overall, N(%)	Poverty Condition (Based on LPL)		p-value	Poverty Condition (Based on UPL)		p-value
		HHs Below LPL, n(%)	Otherwise, n(%)		HHs Below UPL, n(%)	Otherwise, n(%)	
Cultivable Land (in Decimal)							
No land	1589(67.9)	452(78.2)	1137(64.5)		951(73.7)	638(60.8)	
1-15	157(6.7)	47(8.1)	110(6.2)	<0.001	99(7.7)	58(5.5)	<0.001
16-50	191(8.2)	37(6.4)	154(8.7)		99(7.7)	92(8.8)	
50+	403(17.2)	42(7.3)	361(20.5)		141(10.9)	262(25.0)	
Electricity Connection							
No	451(19.3)	151(26.1)	300(17.0)	<0.001	265(20.5)	186(17.7)	0.085
Yes	1889(80.7)	427(73.9)	1462(83.0)		1025(79.5)	864(82.3)	
Livestock Ownership							
No	1359(58.1)	317(54.8)	1042(59.1)	0.070	738(57.2)	621(59.1)	0.346
Yes	981(41.9)	261(45.2)	720(40.9)		552(42.8)	429(40.9)	
		Average Values (SD)			**Average Values (SD)**		
Age (in Year)	42.8(11.5)	41.8(10.0)	43.2(11.9)	0.014	41.9(10.5)	43.9(1.4)	<0.001

Note: HHs=Households, SD=Standard Deviation
Source: Author's calculation from survey data 2019.

From Table 2, approximately 25% of the households lie below the LPL, and 55% of them lie below the UPL. The proportion of households below the LPL and UPL was found to be highest (54.7% below LPL and 53.3% below UPL) in the Sunamganj district and lowest (0.7% below LPL and 1.8% below UPL) in the Sylhet district. The proportion of households below the poverty level was higher for the borrower (60.2% below LPL and 63.7% below UPL) than the non-borrower (39.8% below LPL and 36.3% below UPL) households, which might imply that the borrower households received micro-credits due to their poverty conditions.

The average household size was found to be 4.9±1.5, and a larger portion of households (97.8% below LPL and 93.2% below UPL) that have four or more members lie below the poverty line. The percentage of household heads below the LPL was highest among married heads (95.0%) and lowest among unmarried heads (0.9%). Moreover, 4.2% of the widowed/divorced household heads lie below the LPL. About 62% of the households that lie below the LPL and 54% that lie below the UPL are not members of NGOs.

The percentage of households that lie below the poverty line was highest (71.5% below LPL and 73.5% below UPL) in the income group 10,000-20,000 Taka and lowest (0.7% below LPL and 2.0% below UPL) in the income group 30,000 Taka or more. The percentage of households below the poverty line was highest (78.2% below LPL and 73.7% below UPL) for those with no land and lowest (6.4% below LPL and 7.7% below UPL) for those with land 16-50 decimals. The average age of the heads of households below the LPL (41.8±10.0 years) was close to that of the heads of households below the UPL (41.9±10.5 years).

Almost all the covariates were highly significantly ($p<0.001$) related to the poverty conditions based on LPL. However, marital status and livestock ownership were found to be significant, with a p-value of 0.004 and 0.070, respectively. Regarding UPL, all factors except marital status and ownership of livestock were found to be significantly (p<0.05) related to the poverty conditions. However, access to electricity was significant, with a p-value = 0.085.

4.2 Identification of the determinants of poverty using the MLR model

Table 3: Determinants of Households' Poverty Conditions Using the MLR Model

Determinants	**Results of MLR Model Based on LPL**			**Results of MLR Model Based on UPL**		
	Beta	**OR(95% CI)**	**p-value**	**Beta**	**OR(95% CI)**	**p-value**
District						
Sunamganj: Ref.	-	1.00	-	-	1.00	-
Sylhet	-1.34	0.26(0.09-0.78)	0.016	-1.01	0.37(0.19-0.69)	0.002
Habiganj	-0.02	0.98(0.69-1.39)	0.914	-0.08	0.93(0.66-1.31)	0.672
Netrokona	-0.47	0.63(0.42-0.93)	0.021	-0.24	0.79(0.56-1.12)	0.185
Kishoreganj	0.14	1.14(0.79-1.65)	0.467	-0.61	0.55(0.38-0.78)	0.001
Micro-Credit Status						
Non-Borrower: Ref.	-	1.00	-	-	1.00	-
Borrower	-0.64	0.53(0.39-0.71)	<0.001	-1.11	0.33(0.24-0.46)	<0.001
Household Size						
<4: Ref.	-	1.00	-	-	1.00	-<0.001
4 or more	2.22	9.19(4.75-17.82)	<0.001	1.38	3.99(2.72-5.85)	
Age (in Year)	-0.02	0.98(0.97-0.99)	0.003	-0.02	0.98(0.97-0.99)	<0.001
Gender						
Male: Ref.	-	1.00	-	-	1.00	-
Female	0.04	1.04(0.54-1.99)	0.914	-0.01	0.99(0.54-1.82)	0.980
Occupation						
Farming: Ref.	-	1.00	-	-	1.00	-
Day laborer	0.35	1.42(0.97-2.09)	0.070	0.28	1.32(0.92-1.90)	0.137
Off-farm activities	0.33	1.39(0.91-2.15)	0.131	0.34	1.41(0.90-2.19)	0.135
Service/Business	-0.06	0.94(0.62-1.43)	0.770	-0.28	0.76(0.52-1.11)	0.153
Household work	-0.54	0.59(0.28-1.23)	0.158	-0.41	0.66(0.33-1.31)	0.233
Others	0.32	1.37(0.78-2.43)	0.277	0.21	1.24(0.70-2.18)	0.462
Marital Status						
Married: Ref.	-	1.00	-	-	1.00	-
Unmarried	-1.03	0.36(0.12-1.03)	0.058	0.44	1.56(0.72-3.38)	0.263
Widowed/Divorced	0.33	1.39(0.72-2.69)	0.329	0.87	2.39(1.31-4.36)	0.005
NGO Membership						
No: Ref.	-	1.00	-	-	1.00	-
Yes	-0.52	0.60(0.45-0.79)	<0.001	-0.25	0.78(0.59-1.02)	0.068
Per Capita Income (in Taka) Per Year						
≤10000: Ref.	-	1.00	-	-	1.00	-
10000-20000	-1.79	0.17(0.11-0.25)	<0.001	-0.77	0.47(0.28-0.78)	0.003
20000-30000	-4.99	0.01(0.00-0.01)	<0.001	-3.58	0.03(0.02-0.05)	<0.001
30000 or more	-5.29	0.01(0.00-0.02)	<0.001	-4.71	0.01(0.01-0.02)	<0.001

Determinants	Results of MLR Model Based on LPL			Results of MLR Model Based on UPL		
	Beta	OR(95% CI)	p-value	Beta	OR(95% CI)	p-value
Cultivable Land (in Decimal)						
No land: Ref.	-	1.00	-	-	1.00	-
1-15	0.13	1.13(0.72-1.78)	0.587	0.19	1.23(0.77-1.92)	0.400
16-50	-0.22	0.80(0.49-1.29)	0.360	0.21	1.23(0.80-1.89)	0.346
50+	-0.96	0.38(0.25-0.59)	<0.001	-0.79	0.46(0.33-0.64)	<0.001
Electricity Connection						
No: Ref.	-	1.00	-	-	1.00	-
Yes	-0.77	0.46(0.34-0.64)	<0.001	-0.37	0.69(0.51-0.95)	0.023
Livestock Ownership						
No: Ref.	-	1.00	-	-	1.00	-
Yes	0.50	1.65(1.27-2.14)	<0.001	0.27	1.31(1.03-1.68)	0.029
Constant	1.07	2.92	0.036	3.34	28.33	<0.001
	Hosmer and Lemeshow Chi-square=12.461, p-value=0.132			Hosmer and Lemeshow Chi-square=9.130, p-value=0.331		

Note: OR=odds ratio, CI=confidence interval
Source: Author’s calculation from survey data 2019.

Table 3 shows the results of the MLR models to identify the determinants of poverty conditions LPL and UPL. The p-value of the Hosmer-Lemeshow test suggests that both of the MLR models fit the data well. The estimated MLR model suggested that district, micro-credit status, household size, age, NGO membership, per capita income, cultivable land, electricity connection, and livestock ownership are the significant ($p<0.05$) determinants of poverty level LPL. On the other hand, district, micro-credit status, household size, age, marital status, per capita income, cultivable land, electricity connection, and livestock ownership are the significant ($p<0.05$) determinants of poverty level UPL.

The findings of the study demonstrated that households in the Sylhet district had about 74% and 63% lower risk of lying below the LPL and UPL, respectively, compared to the households in the Sunamganj district. Receiving micro-credit typically eliminates financial constraints over time by increasing income (Rahman, 2007). In this context, this study found that borrower households had about 47% and 67% lower risk of lying below the LPL and UPL, respectively, in comparison to non-borrower households.

Households having four or more members were 9 times and about 4 times more likely to lie below the LPL and UPL, respectively, than households having less than four members. The study included the respondent’s age as a potential determinant in the model and found that for every one-unit increase in the respondent’s age, the likelihood of falling below the LPL and UPL decreased by 2%.

Widowed/divorced household heads had a 2.4 times higher risk of lying below the UPL than married household heads. The NGO member households had a 40% lower risk of lying below the LPL than the non-member households. The risk of lying below the LPL and UPL was, respectively, 83% and about 53% lower for a household having an income of 10,000-20,000 Taka compared to a household having an income of 10,000 Taka or less.

Households with marginally cultivable land (more than 50 decimals) had about 62% and 54% lower risk of lying below the LPL and UPL, respectively, than landless households. Households with access to electricity had about 54% and 31% lower probability of lying below the LPL and UPL, respectively, than households having no electricity access.

4.3 Performance evaluation of machine learning algorithms

The performance of several ML models was evaluated using accuracy, precision, recall, F1-score, mean absolute error (MAE), and root mean square error (RMSE). Also, the CM and AUROC for performance comparison were evaluated.

4.3.1) Accuracy, precision, recall, F1-score, MAE, and RMSE

Table 4: Overall Performance Metrics of ML Models to Predict Poverty Conditions Based on LPL

Models	**Poverty Conditions**	**Precision**	**Recall**	**F1-Score**	**MAE**	**RMSE**	**Accuracy**
SVM	Below LPL	0.71	0.50	0.59	0.18	0.43	0.82
	Otherwise	0.84	0.93	0.88			
Naïve Bayes	Below LPL	0.58	0.55	0.56	0.22	0.47	0.78
	Otherwise	0.84	0.86	0.85			
Logistic Regression	Below LPL	0.64	0.50	0.56	0.20	0.45	0.80
	Otherwise	0.84	0.90	0.87			
KNN, **K=4**	Below LPL	0.70	0.40	0.51	0.19	0.45	0.80
	Otherwise	0.82	0.94	0.88			
Decision Tree	Below LPL	0.61	0.50	0.55	0.21	0.45	0.79
	Otherwise	0.84	0.89	0.86			
Random Forest	Below LPL	0.62	0.43	0.51	0.22	0.47	0.78
	Otherwise	0.82	0.91	0.86			
Outcome variable: Poverty condition (based on LPL)							

Source: Author’s calculation from survey data 2019.

Table 5: Overall Performance Metrics of ML Models to Predict Poverty Conditions Based on UPL

Models	**Poverty Conditions**	**Precision**	**Recall**	**F1-Score**	**MAE**	**RMSE**	**Accuracy**
SVM	Below UPL	0.82	0.84	0.83	0.21	0.46	0.79
	Otherwise	0.76	0.73	0.74			
Naïve Bayes	Below UPL	0.82	0.84	0.83	0.20	0.45	0.80
	Otherwise	0.76	0.75	0.75			
Logistic Regression	Below UPL	0.83	0.85	0.84	0.19	0.44	0.81
	Otherwise	0.78	0.75	0.76			
KNN, **K=5**	Below UPL	0.82	0.80	0.81	0.22	0.47	0.78
	Otherwise	0.73	0.74	0.73			
Decision Tree	Below UPL	0.82	0.84	0.83	0.21	0.46	0.79
	Otherwise	0.76	0.73	0.74			
Random Forest	Below UPL	0.83	0.73	0.78	0.24	0.49	0.76
	Otherwise	0.68	0.79	0.73			
Outcome variable: Poverty condition (based on UPL)							

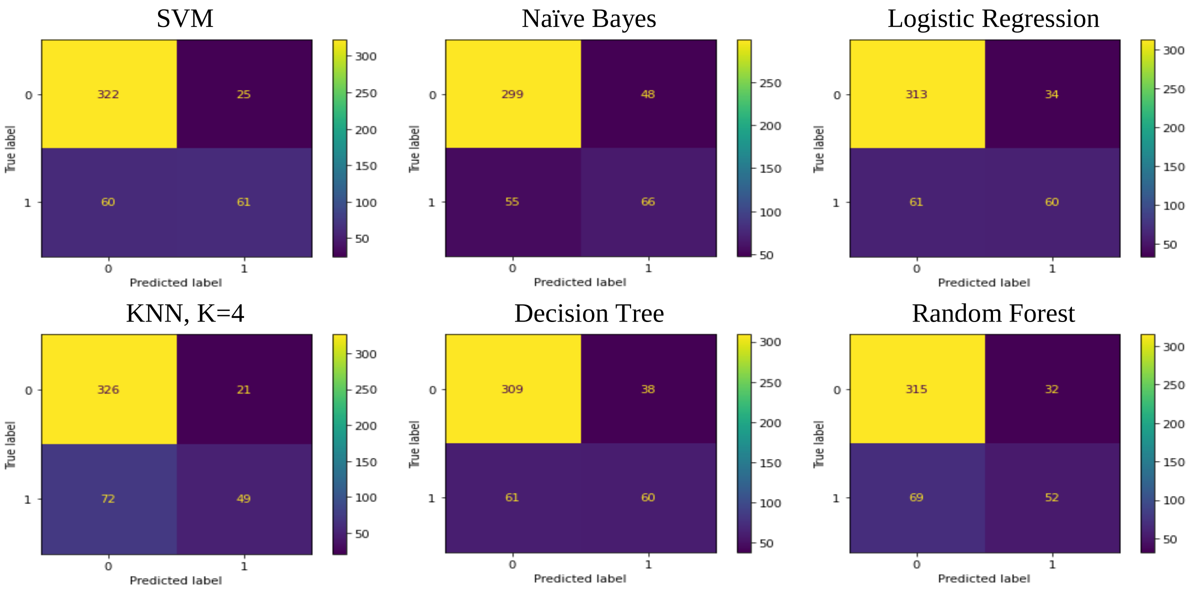
Source: Author’s calculation from survey data 2019.

From Table 4, the SVM has the highest accuracy (82%) and F1-score (59%) in predicting the poverty condition based on LPL. From Table 5, the logistic regression model has the highest accuracy (81%) and F1-score (84%) in predicting the poverty condition based on UPL. These two models also have the lowest MAE and RMSE compared to other models.

Therefore, the SVM and logistic regression are the best models to predict poverty conditions based on LPL and UPL, respectively, in terms of accuracy and F1-score.

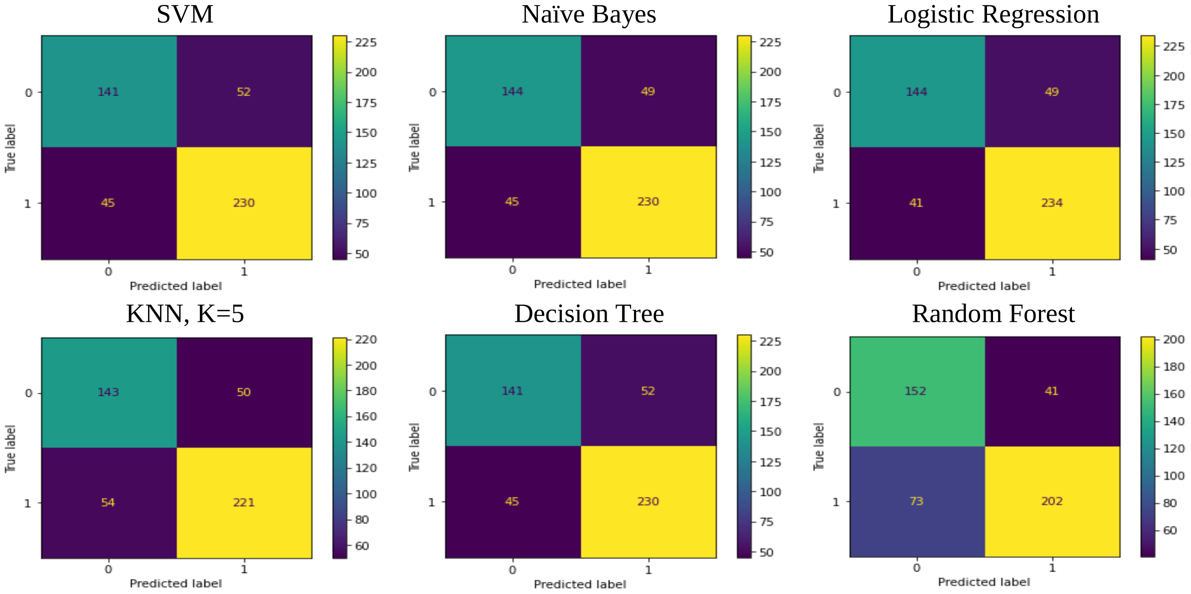
4.3.2) Confusion matrix

Figure 2: Confusion Matrix of ML Models to Predict Poverty Conditions Based on LPL



Source: Author’s summarization.

Figure 3: Confusion Matrix of ML Models to Predict Poverty Conditions Based on UPL



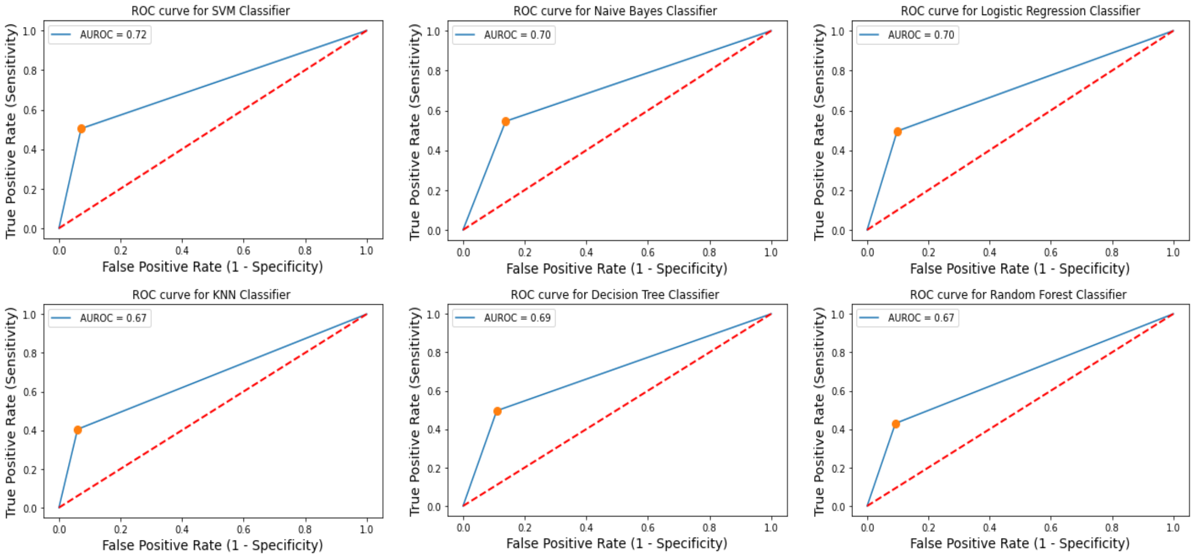
Source: Author’s summarization.

To determine the performance of ML models, we look into the TP, TN, FP, and FN of each algorithm. From Figure 2, the SVM predicts 82% correct classification, which is the highest among all other models, and 18% incorrect classification, which is the lowest among all other models. Again, from Figure 3, the logistic regression model predicts 81% correct classification, which is the highest among all other models, and 19% incorrect classification, which is the lowest among all other models.

Therefore, the confusion matrix ensures that the SVM and logistic regression are the best models to predict the poverty level LPL and UPL, respectively.

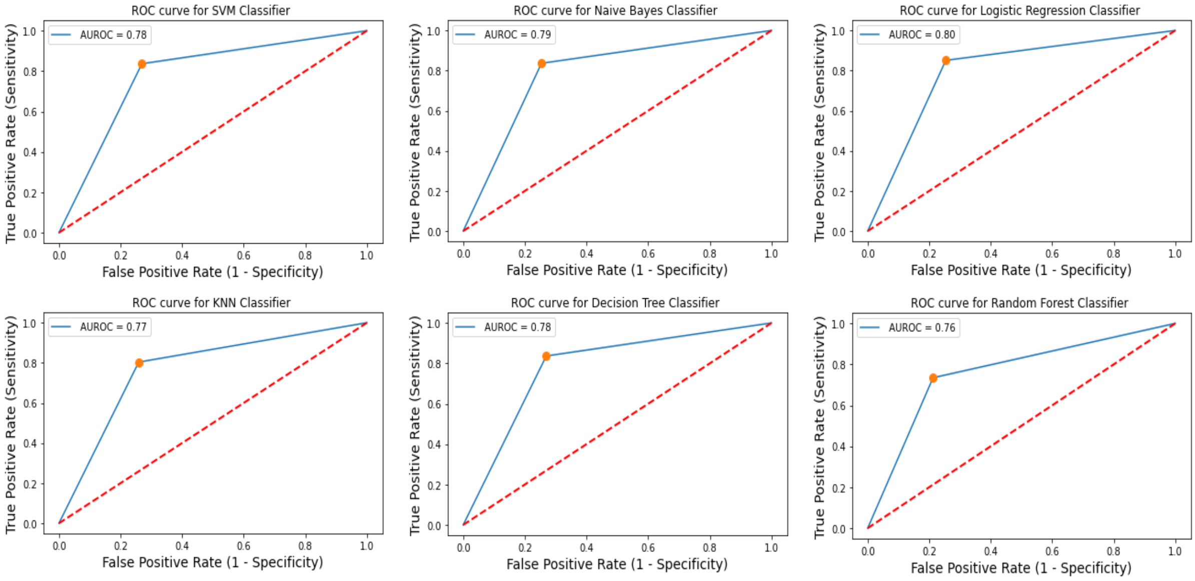
4.3.3) Receiver operating characteristics system

Figure 4: ROC Curve of ML Models to Predict Poverty Conditions Based on LPL



Source: Author's summarization.

Figure 5: ROC Curve of ML Models to Predict Poverty Conditions Based on UPL



Source: Author's summarization.

To validate further the performance of several algorithms, the ROC curve and the AUROC have been evaluated. From Figure 4, the ROC curve of SVM, and from Figure 5, the ROC curve of the logistic regression model appears to be better (closer to the upper left corner) compared to the other models. Moreover, the SVM has the highest AUROC of 72% (Figure 4), and the logistic regression model has the highest AUROC of 80% (Figure 5), indicating the excellent discrimination of the model (Yang et al., 2017).

Therefore, the SVM and logistic regression are the best models to predict the poverty level LPL and UPL, respectively, in terms of AUROC.

4.3.4) Stratified 5-fold cross-validation

Table 6: Stratified 5-Fold Cross-Validation of ML Models to Predict Poverty Conditions Based on LPL

Models	**Accuracy**					
	K-1	**K-2**	**K-3**	**K-4**	**K-5**	**Average**
SVM	0.81	0.85	0.85	0.78	0.81	0.82
Naïve Bayes	0.79	0.80	0.80	0.78	0.82	0.80
Logistic Regression	0.81	0.83	0.83	0.78	0.81	0.81
KNN, **K=4**	0.81	0.83	0.83	0.78	0.80	0.81
Decision Tree	0.81	0.79	0.83	0.77	0.81	0.80
Random Forest	0.80	0.79	0.82	0.75	0.79	0.79
Outcome variable: Poverty condition (based on LPL)						

Source: Author’s calculation from survey data 2019.

Table 7: Stratified 5-Fold Cross-Validation of ML Models to Predict Poverty Conditions Based on UPL

Models	**Accuracy**					
	K-1	**K-2**	**K-3**	**K-4**	**K-5**	**Average**
SVM	0.81	0.80	0.83	0.82	0.81	0.81
Naïve Bayes	0.81	0.81	0.83	0.85	0.80	0.82
Logistic Regression	0.81	0.83	0.83	0.83	0.81	0.82
KNN, **K=5**	0.79	0.77	0.81	0.78	0.78	0.78
Decision Tree	0.80	0.80	0.83	0.84	0.82	0.82
Random Forest	0.76	0.79	0.77	0.78	0.77	0.77
Outcome variable: Poverty condition (based on UPL)						

Source: Author’s calculation from survey data 2019.

From Table 6, the SVM still outperforms the other five models with the highest average accuracy (82%) in predicting the poverty level LPL. From Table 7, the logistic regression model outperforms the other five models with the highest average accuracy (82%) in predicting the poverty level UPL. Although the Naïve Bayes and decision tree classifiers have the same average accuracy as the logistic regression classifier, their other performance metrics are worse than the logistic regression classifier in predicting the poverty level UPL.

Therefore, SVM and logistic regression are the best models to predict the poverty level LPL and UPL, respectively.

5. Discussion

Poverty is one of the main obstacles to the socioeconomic development of a country or a society. In order to end poverty in a particular region, we need to predict the level of poverty of that regional household. This study intended to identify the determinants of poverty and select the best machine learning model to predict the poverty conditions of the north-eastern wetland region of Bangladesh.

Machine learning algorithms are widely used in many fields, including data mining, to predict outcomes, identify patterns, and extract meaningful insights from large datasets. As a result, this study considered several supervised type machine learning algorithms such as support vector machine, Naïve Bayes, logistic regression, K-nearest neighbor, decision tree, and random forest to predict the poverty conditions from previous research (Sani et al., 2018; Shen, 2021; Talingdan, 2019), and their performances were measured by using accuracy, precision, recall, F1-score, CM, and AUROC. This study also used the MLR model to extract features for ML models or to identify the factors that determine poverty situations.

The study found several individual-level and household-level factors, such as district, micro-credit status, household size, age, NGO membership, marital status, per capita income, cultivable land, electricity connection, livestock ownership, etc., that determine the wetland people’s poverty conditions. According to findings, households in urban areas (Sylhet and Netrokona) are less likely to be poor than those in typical rural areas (Sunamganj). Similar findings were reported by Kazal et al. (2017).

Borrower and NGO member households had a lower risk of being poor than the non-borrower and non-NGO member households. Thus micro-credit and NGO programs have the efficacy to eradicate poverty in the short run. The likelihood of being poor increases with the increase in household size. The reason for the findings may be that the opportunity for per capita food intake tends to decline with the increase in the size of households. The likelihood of being poor decreases with the increase in the respondent’s age. The apparent explanation may be that as people age, they tend to acquire more assets.

Widowed/divorced household heads had a higher risk of being poor than married household heads. One obvious reason may be that widowed women are prone to losing rights of access to properties such as land, housing, etc., that they enjoyed during the lifetime of their husbands (Doss et al., 2012). Such alienation from property is linked to poverty (Carter & Barrett, 2006). The risk of being poor decreases as the income and cultivable land of the wetland people increases.

The availability of electricity is effective in reducing poverty among households in the wetland region. The outcome agreed with a study carried out in Bangladesh by Imam et al. (2018). The fact is that having access to electricity allows a variety of activities due to its direct or indirect links to employment and high-return industries. It is recommended that the wetland people should be given attention to these determinants of poverty.

The study also found that support vector machine and logistic regression are the best models to predict the poverty level LPL and UPL, respectively based on the features extracted from the MLR features selection technique. Because the models have the highest accuracy, F1-score, and AUROC, and the lowest MAE and RMSE among all the models considered in this study.

6. Conclusions and Policy Recommendations

The study's findings provide valuable insights into the effective use of machine learning algorithms in precisely identifying the target poor populations and identifying the determinants of poverty reduction of these impoverished populations living in the wetland region of Bangladesh. This study identified several factors affecting poverty in the wetland region of Bangladesh and two machine learning models to predict poverty conditions. The findings of the study conclude that living in urban areas, receiving micro-credit, having a small number of family members, being a member of NGOs, having high income and cultivable land, and having electricity access in households may be protective towards reducing poverty in the study area. This study has made the following policy recommendations based on the findings.

The government can develop the condition of poor households in rural areas by creating facilities for income-generation activities (IGAs). By employing resources like wetlands, rich soil, and biodiversity, a "nature-based solution" strategy can improve opportunities for IGAs in the wetland region. Seasonal fish farming and climate-smart agricultural practices like raising ducks and cattle, as well as floating vegetable gardens, are examples of potential IGAs. Creating an eco-friendly travel sector has potential as well. It is necessary to simplify the government micro-credit program and extend it with more lenient terms and conditions in the wetland region. The high interest rate and risk of asset depletion associated with micro-credits from unofficial sources should be eliminated. It is necessary to adjust the interest rates of the current micro-credit programs from non-governmental (MFI, NGO, and insurance) sources for the wetland area.

The involvement of various NGOs might be beneficial for the wetland people by achieving skills development training to handle any IGAs with competence. The family planning program may be strengthened in the wetland area to maintain the optimum family size. Landless people in wetland areas may be encouraged to participate in sharecropping. The government can ensure access to electricity in the wetland region due to its direct and indirect links with IGAs. The implementation of all these policies may help to achieve SDG goal 1.

References

Acharya, K. P., Khanal, S. P., & Chhetry, D. (2022). Factors affecting poverty in Nepal-A binary logistic regression model study. *Pertanika Journal Social Science and Humanities*, *30*(2), 641-663.

Achia, T. N. O., Wangombe, A., & Khadioli, N. (2010). A logistic regression model to identify key determinants of poverty using demographic and health survey data. *European Journal of Social Sciences*, *13*(1), 38-45.

Alsharkawi, A., Al-Fetyani, M., Dawas, M., Saadeh, H., & Alyaman, M. (2021). Poverty classification using machine learning: The case of Jordan. *Sustainability*, *13*(3), 2-16.

Bangladesh Bureau of Statistics. (2023). *Key findings: Household income and expenditure survey-2022*. Statistics and Informatics Division, Ministry of Planning, Government of the People's Republic of Bangladesh, Dhaka, Bangladesh. Retrieved from https://bbs.portal.gov.bd/sites/default/files/files/bbs.portal.gov.bd/page/57def76a_aa3c_46e3_9f80_53732eb94a83/2023-04-13-09-35-ee41d2a35dcc47a94a595c88328458f4.pdf

Bangladesh Bureau of Statistics. (2017). *Preliminary report of household income and expenditure survey-2016*. Statistics and Informatics Division, Ministry of Planning, Government of the People's Republic of Bangladesh, Dhaka, Bangladesh. Retrieved from https://bbs.portal.gov.bd/sites/default/files/files/bbs.portal.gov.bd/page/b343a8b4_956b_45ca_872f_4cf9b2f1a6e0/HIES%20Preliminary%20Report%202016.pdf.

Biyase, M., & Zwane, T. (2017). *An empirical analysis of the determinants of poverty and household welfare in South Africa*. Retrieved from Munich Personal RePEc Archive https://mpra.ub.uni-muenchen.de/77085/

Borko, Z. P. (2017). Determinants of poverty in rural households (the case of Damot Gale district in Wolaita Zone, Ethiopia): A household level analysis. *International Journal of African and Asian Studies*, *29*, 68-75.

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

Carter, M. R., & Barrett, C. B. (2009). The economics of poverty traps and persistent poverty: An asset-based approach. *Journal of Development Studies*, *42*(2), 178-99.

Centre for Environmental and Geographic Information Services. (2012). *Master plan of Haor area.* Ministry of Water Resources, Bangladesh Haor and Wetland Development Board, Dhaka, Bangladesh. Retrieved from https://dbhwd.portal.gov.bd/sites/default/files/files/dbhwd.portal.gov.bd/publications/baf5341d_f248_4e19_8e6d_e7ab44f7ab65/Haor%20Master%20Plan%20Volume%201.pdf

Chowdhury, A. (2014). *Factors affecting productivity and efficiency of rice production in haor area in Bangladesh: Likely impact on food security* (Master's Thesis, Bangladesh Agricultural University, Mymensingh, Bangladesh). Retrieved from https://catalog.ihsn.org/citations/42191

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning*, *20*(3), 273-297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21-27

Doss, C., Truong, M., Nabanoga, G., & Namaalwa, J. (2012). Women, marriage and asset inheritance in Uganda. *Development Policy Review*, *30*(5), 597-616.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York, USA: John Wiley and Sons. Retrieved from https://www.semanticscholar.org/paper/Pattern-classification-and-scene-analysis-Duda Hart/b07ce649d6f6eb636872527104b0209d3edc8188

Fix, E., & Hodges, J. L. (1951). *Discriminatory analysis: Nonparametric discrimination, small sample performance*. USA: Air University, USAF School of Aviation Medicine, 1952.

Hasan, M. R., & Hossain, M. Z. (2024). Factors affecting formal micro-credits in the wetland regions of Bangladesh: A discriminant analysis. *International Journal of Statistical Sciences*, *24*(2), 85-96.

Hashemi, S. M., Schuler, S. R., & Riley, A. P. (1996). Rural credit programs and women's empowerment in Bangladesh. *World Development*, *24*(4), 635-653.

Imam, M. F., Islam, M. A., & Hossain, M. J. (2018). Factors affecting poverty in rural Bangladesh: An analysis using multilevel modelling. *Journal of the Bangladesh Agricultural University*, *16*(1), 123-130.

Kambuya, P. (2020). Better model selection for poverty targeting through machine learning: A case study in Thailand. *Thailand and the World Economy*, *38*(1), 91-116.

Kazal, M. M. H., Rahman, S., & Hossain, M. Z. (2017). Poverty profiles and coping strategies of the haor (ox-bow lake) households in Bangladesh. *Journal of Poverty Alleviation and International Development*, *8*(1), 167-191.

Khondker, B. H., & Mahzab, M. M. (2015). *Lagging districts development: Background study paper for preparation of the seventh five-year plan*. Retrieved from https://www.researchgate.net/publication/332567169_Lagging_Districts_Development_Background_Study_Paper_for_Preparation_of_the_Seventh_Five-Year_Plan

Kim, J. Y. (2021). Using machine learning to predict poverty status in Costa Rican households. *SSRN Electronic Journal*, 1-13. Retrieved from https://doi.org/10.2139/ssrn.3971979

Korankye, A. A. (2014). Causes of poverty in Africa: A review of literature. *American International Journal of Social Science*, *3*(7), 147-153.

Langley, P., Iba, W., & Thompson, K. (1992, January). An analysis of Bayesian classifiers. *Proceedings of the 10th National Conference on Artificial Intelligence*, San Jose: AAAI Press, 223-228. Retrieved from https://cdn.aaai.org/AAAI/1992/AAAI92-035.pdf

Li, Q., Yu, S., Échevin, D., & Fan, M. (2022). Is poverty predictable with machine learning? A study of DHS data from Kyrgyzstan. *Socio-Economic Planning Sciences*, *81*.

Min, P. P., Gan, Y. W., Hamzah, S. N. B., Ong, T. S., & Sayeed, M. S. (2022). Poverty prediction using machine learning approach. *Journal of Southwest Jiaotong University*, *57*(1), 136-146.

Ministry of Law. (2013). *Bangladesh Water Act, 2013.* Justice and Parliamentary Affairs, Legislative and Parliamentary Affairs Division, Government of the People's Republic of Bangladesh. Retrieved from http://oldweb.lged.gov.bd/uploadeddocument/UnitPublication/1/840/Water%20Act%202013%20(English).pdf

Mohamud, J. H., & Gerek, O. N. (2019, April). *Poverty level characterization via feature selection and machine learning*. IEEE 2019 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, 1-4. Retrieved from http://dx.doi.org/10.1109/SIU.2019.8806548

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2012). *Generalized linear models with applications in engineering and the sciences*. New York, USA: John Wiley and Sons.

Ogwumike, F. O., & Akinnibosun, M. K. (2013). Determinants of poverty among farming households in Nigeria. *Mediterranean Journal of Social Sciences*, *4*(2), 365-373.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106.

Rahman, S. (2007). *The impact of microcredit on poverty and women's empowerment: A case study of Bangladesh* (Doctoral Thesis, Western Sydney University, Sydney, Australia). Retrieved from http://handle.uws.edu.au:8081/1959.7/36990

Rhoumah, A. (2016). Determinants of factors that affect poverty among coastal fishermen community in Malaysia. *IOSR Journal of Economics and Finance*, *7*(3), 9-13.

Sani, N. S., Rahman, M. A., Bakar, A. A., Sahran, S., & Sarim, H. M. (2018). Machine learning approach for bottom 40 percent households (B40) poverty classification. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(4-2), 1698-1705.

Santa, G. M., & Ruiz, L. C. M. (2023). Predicting multidimensional poverty with machine learning algorithms: An open data source approach using spatial data. *Social Sciences*, *12*(5), 2-21.

Santoso, S., & Irwan, M. I. (2016). Classification of poverty levels using k-nearest neighbor and learning vector quantization methods. *International Journal of Computing Science and Applied Mathematics*, *2*(1), 8-13.

Shen, T., Zhan, Z., Jin, L., Huang, F., & Xu, H. (2021, June). *Research on method of identifying poor families based on machine learning*. IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 10-13. Retrieved from https://doi.org/10.1109/IMCEC51613.2021.9482142

Sheng, W., Yumei, S. (2021). *Prediction poverty levels of college students using a machine learning model*. Retrieved from Research Square https://doi.org/10.21203/rs.3.rs-919541/v1

Sohnesen, T. P., & Stender, N. (2017). Is random forest a superior methodology for predicting poverty? An empirical assessment. *Poverty & Public Policy*, *9*(1), 118-133.

Spaho, A. (2014). Determinants of poverty in Albania. *Journal of Educational and Social Research*, *4*(2), 157-163.

Talingdan, J. A. (2019, May). *Performance comparison of different classification algorithms for household poverty classification*. IEEE 2019 4th International Conference on Information Systems Engineering, Shanghai, China, 11-15. Retrieved from http://dx.doi.org/10.1109/ICISE.2019.00010

Thoplan, R. (2014). Random forests for poverty classification. *International Journal of Sciences: Basic and Applied Research*, *17*(2), 252-259.

Wang, S., Zhao, Y., & Zhao, Y. (2020). Costa Rican poverty level prediction. *IETI Transactions on Social Sciences and Humanities*, *7*(2020/05), 171-176.

Wong, G. (2022). Poverty prediction and the identification of discriminative features on household data from Cambodia. *TechRxiv*, 1-6.

Yang, S., & Berdine, G. (2017). The receiver operating characteristic (ROC) curve. *The Southwest Respiratory and Critical Care Chronicles*, *5*(19), 34-36.

Zixi, H. (2021, March). *Poverty prediction through machine learning*. 2nd International Conference on E-Commerce and Internet Technology, Hangzhou, China, 314-324. Retrieved from https://doi.org/10.1109/ECIT52743.2021.00073