STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–12.
Published online in International Academic Press (www.IAPress.org)

Poverty prediction using machine learning models: Insights from HICES survey in Egypt

Israa Lewaaelhamd *, Maged George Iskander

Faculty of Business Administration, Economics and Political Science, Department of Business Administration, The British University in Egypt, Cairo, Egypt

Abstract This study focuses on the poverty problem in Egypt. Data from household expenditure and income surveys is used to determine the poverty status of Egyptian households. Nevertheless, conducting these kinds of surveys is challenging, costly, and time-consuming. This procedure might be revolutionized by machine learning. This work contributes to the field by utilizing machine learning techniques to evaluate and track the poverty levels of Egyptian households. This method brings poverty detection closer to real-time, and lower costs, and accuracy. A significant portion of this work involves managing unbalanced data and preparing data. Eleven machine learning classification models are applied. The classification algorithms of the Gradient Boosting Machine and support vector machine have achieved the best performance. The final machine learning classification model could transform efforts to track and target poverty across the country. This work demonstrates how powerful and versatile machine learning can be, and hence, it promotes adoption across many domains in both the private sector and government.

Keywords Sustainability; Data Analytics; Machine Learning Algorithms; Classification Problem, HICES, Poverty.

AMS 2010 subject classifications 65C20, 62M20

DOI: 10.19139/soic-2310-5070-2082

1. Introduction

The first goal of the Sustainable Development Goals is to eradicate poverty in all of its manifestations globally by 2030. The most recent data from the World Bank, however, suggests that most nations will continue suffering from the COVID-19 pandemic's consequences until 2030. Given these circumstances, it is now unrealistic to expect to reduce the absolute world poverty not to exceed 3% by 2030—a goal that was in jeopardy even before to the crisis—without taking prompt, meaningful, and substantial political action.

The Egyptian government makes extensive use of the poverty rate to track and assess the issue of poverty. The population percent of those who live below the absolute poverty level is known as the poverty rate. The National Household Income Expenditure and Consumption Survey (HICES) provides official poverty statistics. Data from HICES survey is used to determine how impoverished Egyptian households are. Nevertheless, conducting these kinds of surveys is challenging, costly, and time-consuming. Even worse, by the time data are gathered and examined, they are frequently outdated, which increases the likelihood that decision-makers would base their choices on outdated information. Machine learning has the potential to completely alter the game by bringing

*Correspondence to: Israa Lewaaelhamd, E-mail: israalewaa@feps.edu.eg or israa.lewaa@bue.edu.eg

ISSN 2310-5070 (online) ISSN 2311-004X (print)

poverty detection and assessment closer to real-time and at a lower cost. To obtain precise national-level data on the economic standing of households within a nation, national level surveys gathered from well-representative samples are usually costly and time-consuming, which is why there aren't many of them. It is now evident that using surveys as a direct method of assessing poverty is challenging, expensive, and time-consuming. It typically takes around two years to determine the poverty line and produce the required analysis following the completion of the national HICES survey. Therefore, a straightforward, precise, and reasonably priced instrument to evaluate and track Egyptian households' poverty status is desperately needed. However, this is a topic that hasn't gotten much coverage in the literature when it comes to Egypt.

Machine learning is a field of computer science that aims to develop algorithms that can learn from data to make predictions. Machine learning is expanding quickly and has significantly influenced several aspects of computer science as well as other fields [1, 2]. Creating algorithms that can learn from data and make predictions based on that data is the goal of machine learning. For the purpose of resolving classification or regression issues, a variety of machine learning techniques are available, such as decision trees, support vector machines, lasso regression, and ridge regression [3]. To prevent overfitting, the linear regression procedure known as "ridge regression" incorporates a penalty component into the objective function [4]. To promote sparse solutions, Lasso regression is a linear regression approach that includes a penalty term in the objective function [5]. Recursively dividing the data according to the input attributes is the function of decision trees, a kind of non-linear regression model [6]. A regression model known as support vector regression makes predictions by utilizing support vector machines. Conversely, a linear connection between the input characteristics and the output value is assumed in the linear regression model, which is a straightforward regression model [7]. On the other hand, problems with binary classification are the scope of logistic regression [8]. The performance of machine learning models is assessed using several measures, such as R-squared, mean absolute error, and mean squared error. The average absolute difference between the real and predicted values is known as the mean absolute error [9]. This work contributes to predict the poverty status of Egyptian households using several machine learning algorithms. This method considers each household income and expenditure survey and takes poverty detection closer to real-time, with lower costs, and accuracy. In the end, our model is being able to classify the household as a non-poor or poor household.

This paper is organized as follows: Section 2 presents a literature review. Methods and materials used in the study, data preprocessing and the description of machine learning algorithms analysis are described in Section 3. Section 4 shows the results of different machine learning approaches that have been used. The limitations of the study are shown in Section 5. Section 6 discusses the conclusion and future research.

2. Literature Review

A survey of academic papers on poverty is carried out to broaden the analysis and create a solid comparison framework with pertinent journals. The goal is to have a thorough grasp of how machine learning algorithms function while tackling issues related to poverty. This study extensively assesses the advantages and disadvantages of several machine learning algorithms in addition to looking for identifying trends. It is anticipated that this comparison will offer a comprehensive picture of how these algorithms may be used in Egypt to address complicated problems like poverty. [10] applied machine learning methods to classify the poverty status in Jordanian households. The authors used approaches including oversampling, undersampling, SMOTE, and class weights to successfully solve difficulties like class imbalance in the dataset between poor to non-poor households. Light-GBM and Bagged Decision Trees showed the best performance among all machine learning approaches.

[11]used the National Poverty Data for Malaysia to investigate how well k-Nearest Neighbors, Decision Trees, and Naïve Bayes classified the Malaysian households. Their study emphasized the importance of normalization, sample techniques, feature selection, and data preparation. The Synthetic Minority Oversampling Technique (SMOTE) was employed to correct the class disparity. Different combinations of parameters, such as discretization for Naïve Bayes, confidence factor, minimum number of objects for Decision Tree, k-value, and distance function for k-Nearest Neighbours, were used to tune each classifier. Their findings showed that the Decision Tree model's has the best performance among other classifiers. [12] used machine learning models to forecast poverty levels. The

author combined the data from Poverty Probability Index and Oxford Poverty & Human Development Initiative. Data analysis was used to gain a better understanding of the connections between several factors affecting the chance of poverty. Several models were considered including gradient boosting, random forest, decision trees, and linear regression to determine which machine learning model was best for classifying and predicting poverty.

[13]considered the Pakistani household poverty factors. Their study employed logistic regression analysis to examine the personal attributes of the head of the family as well as general household variables as predictors. Based on the monthly expenditures per adult family, the households were divided into quartiles and the lowest quartile was designated as poor, while the remaining three were designated as non-poor. Odd ratios were used to explain the results, and the Wald test was used to determine the significance of the coefficients. The results of the investigation showed a correlation between a higher education level and a lower probability of poverty. Furthermore, the chance of poverty was significantly decreased by the availability of remittances.

[14]used multinomial and ordinal logistic regression models to examine the variables impacting Poland's poverty rate. Based on family income percentages about the poverty level, three states of poverty were distinguished: below near poverty, near poverty, and above near poverty. Analysis was done using information gathered from biennial household surveys conducted from 2000 to 2015. The ordinal logistic regression model failed to fulfil the premise of parallel lines, leading the study to conclude that the multinomial logit model was a better fit for predicting poverty states. Notably, the most important variables affecting the degree of poverty were found to be education, place of residence, labor force participation, and socioeconomic group. [15] suggests a method for evaluating county-level poverty that combines GIS data with NL remote sensing data using machine learning models. As seen by their findings, the random forest model outperformed the other models in terms of accuracy (R2 = 0.928, MAE = 0.030, RMSE = 0.037). The predictions made by [16] using LASSO logit, random forest, and support vector machines demonstrate high accuracy performance in predicting poverty status; however, the effectiveness of each approach differs depending on the sample used. But in this case, the linear support vector machine beat the polynomial method, indicating that nonlinearity isn't a significant feature of the available data. Complementary information on important variables is provided by the results of random forest and LASSOlogit, which aids in understanding the fundamental workings of the subjective poverty prediction process.

[17]investigates the validity of current methods for model validation through a series of design-based simulation experiments. Their research revealed that the validation process, which is frequently employed in machine learning techniques, might provide inaccurate results for selecting the optimal set of estimates across various techniques and circumstances, which can be deceptive in terms of model evaluation. However, in order to forecast multidimensional poverty both before and after the pandemic, [18] used Multidimensional Poverty Index Data from the Oxford Poverty and Human Development Initiative for the years 2019 and 2021. To address research concerns regarding poverty, the study employs a number of data analytic approaches, including feature correlation and selection, as well as graphical visualizations. To predict poverty across four datasets on a national and subnational scale, a variety of machine learning techniques have been used, including Multiple Linear Regression, Decision Tree Regressor, Random Forest Regressor, XGBoost, AdaBoost, Gradient Boosting, Linear Support Vector Regressor (SVR), Ridge Regression, Lasso Regression, Elastic Net Regression, and K-Nearest Neighbour Regression algorithm. In their analysis, the Ridge Regression model exhibits the greatest R2 score and the best performance.

3. Materials and Methods

3.1. Approach

This paper tackles the poverty prediction problem using popular machine learning methods. It is important to test the performance of existing algorithms, as this research is the first study of this type to be performed in Egypt. The investigation starts with a review and analysis of the current dataset, then addresses issues raised by the data and offers potential solutions. Following processing, the data are supplied to eleven machine learning algorithms. Ultimately, the findings are showcased and deliberated upon. A thorough description of the study's flow is shown in Figure 3.1. The three primary stages of the technique are described in depth in the following sections.

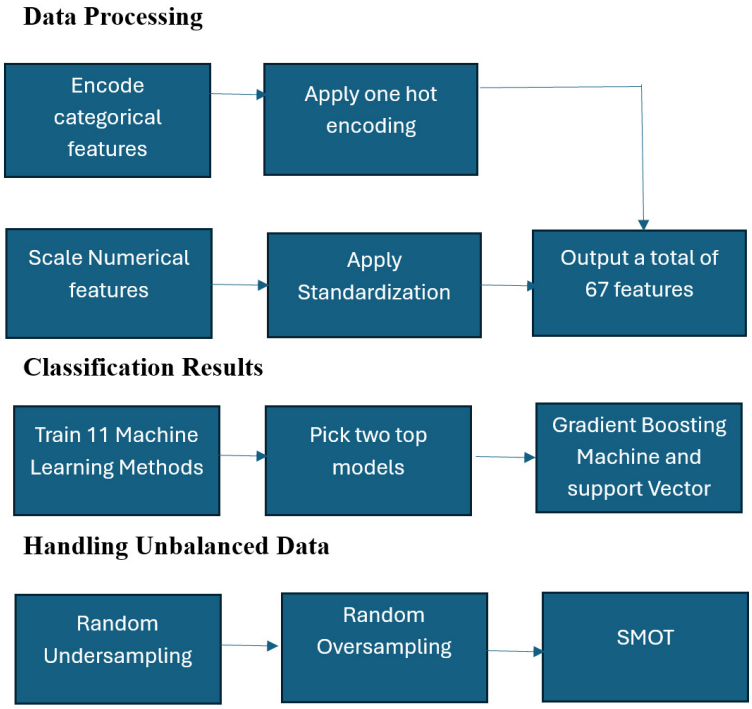

Figure 3.1. Flow chart for used techniques

3.2. Dataset

The data are from the Egyptian Household Income Consumption Expenditure Survey (EHICES), conducted by the Central Agency for Public Mobilization and Statistics in 2017-2018. Data are available on their website. Features used in the study are shown in Table 4 and Table 5 in the appendix. The dataset contains 12,485 household responses with 60 features (51 categorical features and 9 numerical features). These features are described in detail in Table 4 and Table 5.

Our target variable is poverty whether the household is poor or non-poor. All details for conducting and constructing the poverty index is exist on the website of the HICES survey. HICES indicates the poverty index in a binary variable, that is, 0 is for "nonpoor" indicating that the household is relatively well off, and 1 for "poor, indicating that the household is relatively poor. We keep all additional factors that could be connected to poverty after removing the ones that were utilized to create the poverty index. The circumstances of the household are fully described by these factors. We categorize the variables into many groups due to the abundance of explanatory factors at our disposal. A household's demographics are the most frequent element that influences poverty, such as the age, gender, education, marital status of the household head, and region of residence. Materials used in household building, material belongings, and other aspects of assets and quality of life may be a big indicator of one's financial circumstances and level of poverty. Education and health may also be detrimental to poverty. As a result, we separate the explanatory factors into four groups: household situation, quality of life, health status and household demographics. These classifications are indicative in nature; they are neither exhaustive nor ideal.

The household for the HICES were chosen using a two-stage stratified cluster sampling procedure. First, using the sampling frame supplied by the housing and population census, the households are sampled with probabilities proportional with their size. A systematic random sampling technique is used to sample houses for the field test and data collecting at the same time. To provide the indicators to be measured with a high degree of accuracy and efficiency, the survey sample was designed to be representative at the level of the governorates and regions of the

Table 1. Sample Distribution of the governorates

Governorate	Percent	Governorate	Percent
Cairo	9.9	Monofia	4.1
Alexandria	4.2	behira	4.4
Port said	3.9	Ismailia	3.9
Suez	3.6	Giza	5.5
Damietta	3.9	Bani Suef	4
Dakahlia	5.1	Fayoum	3.9
Alsharkia	4.9	Almenia	4
Qalyubia	4	Asuit	3.9
Kafrelshikh	3.9	Sohag	4
Elgharbia	3.9	Qina	3.9
Auxor	4	Aswan	3.9
Border Governerates	3.1		

Republic (urban governorates - urban Upper Egypt - rural Upper Egypt - urban Upper Egypt - rural Upper Egypt - border governorates). Eight households are sampled in the second stage, followed by four households. In the event that any of the initial eight households decline to participate in the survey, the remaining four are sampled as backups. This offers a convincing and scientific method to guarantee that the sample accurately reflects the Egyptian population.

The largest participation rate is from the five governorates in Egypt (9.88% Cairo, 5.46% Bani Soef, 5.06% Dakahlia, 4.91% Alsharkia and 4.19% Alexandriaas shown in Table 1.

3.3. Data Challenges

The information gathered is unbalanced. The approximate ratio of the poor to non-poor, which is 1:3. 3145 households are poor, compared to 9340 non-poor ones. The issue of class imbalance arises when there is an unequal representation of class distributions. The lowest class is the most affected as it was expected that many algorithms for classification learning would have low predicted accuracy.

3.4. Data Preprocessing

Prior to utilizing any machine learning methods to forecast a household's poverty level, the data must be accurate, full, and formatted correctly. Preprocessing the data is a crucial step that might impact the accuracy and overall effectiveness of the predictive model.

3.4.1. Data Cleaning A crucial preprocessing stage in machine learning is data cleaning, which aims to increase the dataset's dependability and quality. Errors and inconsistencies in the data must be corrected, duplicate entries must be removed to avoid bias, and missing values must be imputated or removed. Data cleaning greatly improves the efficiency and performance of machine learning models by making sure the data is correct, consistent, and presented correctly.

The methodology used by the HICES to gather the data is the reason why the collected data has no missing values. As a result, no imputation method is needed. But first, a numeric format needs to be created using the 51 category elements.

3.4.2. Data Transformation Other crucial responsibilities include converting categorical variables into numerical representations, controlling outliers to prevent skewed results. One typical technique is one hot encoding, which uses binary encoding to signal whether a value exists for each value in each category characteristic and creates a dummy variable for each value. The process of converting categorical information into numerical features that may be fed into algorithms for machine learning and deep learning is known as one hot encoding. A binary vector represents a categorical variable in this way: all values in the vector would be 0, with the exception of the ith value, which would reflect the variable's ith category and be represented by 1. The length of the vector is equal to the number of unique categories in the variable. The dataset has 67 characteristics once all category features have been encoded.

A small percentage of the dataset must be kept for testing to assess the prediction model's accuracy (or compare the performance of several models) and offer an objective assessment of the final model's performance. The testing set should be carefully chosen, covering the different classes a model would encounter in an application. There are three distinct ratios for splitting: 70:30, 80:20, and 90:10. 20% of the data in this study are not part of the training set.

3.4.3. Feature Scaling As unscaled data might result in incorrect predictions, having all the characteristics on a comparable scale is crucial before incorporating the data into a machine learning model. Standardization has been used in this investigation. Standardization is a scaling method that centers the data around the mean value with a unit standard deviation.

3.5. Machine Learning Algorithms Implementation and Model architecture

The primary goal is to arrive at the optimal set with the most accurate poverty prediction. In light of this, we first execute the various machine learning models. First, the machine learning model divides the data into two categories: test and train. A collection of examples utilized during the learning process is called training data, and it is used to fit the parameters (like weights). In contrast, the testing data is a set of data that shares the same probability distribution as the training data set but is separate from it. Ultimately, the degree to which the model accurately predicts poverty is evident. The Naïve Bayes algorithm (NB), Decision tree, Random forest, boosting approaches, and logistic regression are the machine learning techniques that have been selected.

Logistic regression [34] is the probability, $P(y = 1 \mid x)$, of a binary classification result using the logistic function

where;

$P(y = 1 \mid x) = \frac{1}{1+e^{-(k.x+b)}}$

k is the parameter weights

x is the features variables

b is the bias term.

Meanwhile, Naive Bayes classifies [33]based on Bayes' Theorem based on independence assumption of independent variables (features). The posterior probability for class S_j is $P(S_j \mid x) = \frac{P(S_j)\prod P(x_i|S_j)}{P(x)}$

where;

$P(S_j)$ is the prior probability of class S_j

$P(x_i|S_j)$ is the likelihood of feature xi given S_j

$P(x)$ is the evidence.

Decision Trees [32]split data based on feature values to create a tree structure. Each split is determined by a feature x_i and a threshold λ. The decision function at a node is: if $x_i \leq \lambda$ then go to left branch, else go to right branch. On the other side, The forecasts from several decision trees are combined using ensemble techniques called random forests [31]. A random subset of the characteristics and data are used to train each tree. The prediction is the average (regression) or majority vote (classification) of all the trees. The final prediction is: $\hat{y} = \frac{1}{R}\sum_{r=1}^{R} f_r(x)$

where;

R is the number of trees

$f_r(x)$ is the prediction of tree r.

In order to classify data, Support Vector Machines (SVM) [35] locate a hyperplane that maximizes the margin between distinct classes. The following equation defines the hyperplane: $k.x + b = 0$

where;

k is the parameter weights

x is the features variables

b is the bias term.

Boosting [33]is a method that builds a powerful predictive model by combining several weak models. It functions by training models one after the other, fixing mistakes created by the earlier models. The ultimate forecast combines all model outputs and assigns weights based on performance by $\hat{y}^{=} sign(\sum_{q=1}^{Q} \alpha_q f_q(x))$

where;

α_q is the weight of the m-th model

$f_q(x)$ is its prediction

4. Results

A type of supervised machine learning called classification divides an observation—in this study, a household survey response—into two groups, the poor and the non-poor, according to the number of predicted characteristics. There are several classification algorithms that may be used to solve the current classification problem, making the process of selecting the best method complex due to the influence of numerous variables. The decision criteria are composed of features connected to data and factors linked to problems. We applied popular machine learning methods in this study.

4.1. Classification Algorithms results

Given the output variable is binary, a selection has been made to employ a collection of widely used supervised classification techniques, which have been implemented utilizing the Python programming language. The chosen procedures of machine learning approaches are; Naïve Bayes algorithm (NB), Decision tree, Random forest, adaptive boosting, and Extreme Gradient Boosting (XGBoost).

Eleven machine learning approaches for classification are used as shown in Table 2. Feature scaling is carried out utilizing standardization, previously discussed to ensure that features are on the same scale. Several classification approaches are evaluated next. The performance of each algorithm is assessed using the f1-score, precision, and recall.

Table 3 displays the performance of the eleven machine learning algorithms. It is evident that the support vector machine and gradient boosting machine algorithms perform better than the other classification methods, irrespective of the scaling strategy employed. Therefore, concentrating on these two machine learning techniques would be beneficial. Table 4 further suggests that the final machine learning model should perform at least 90% in terms of recall, precision, and f1-score. The standardization approach will be taken into consideration when these two algorithms are examined in the upcoming section.

4.2. Handling Unbalanced Data

Problems with classification are often linked to unbalanced data when there is unequal representation of the classes. While machine learning algorithms have demonstrated remarkable success in several real-world applications, the field of learning from unbalanced data is still in its early stages of development. Unbalanced learning is a term used frequently to describe learning from unbalanced data. Techniques for handling unbalanced data may be broadly divided into two classes. By attempting to improve the data's distribution, the first group concentrates on the data itself. The second group, on the other hand, is more concerned with changing the machine learning algorithm than the data. In this paper, both methodologies are applied.

Table 2. Parameters and abbreviations used in classification approaches

Classification algorithms	Parameters
Logistic Regression [19]	Standard parameters
Gradient Boosting Machine [20]	Number of estimators = 100
Stochastic Gradient Descent [21]	Max number of iterations = 1000 Tol = 0.001
Passive Aggressive [22]	Standard parameters
k-Nearest Neighbors [23]	Standard parameters
Decision Tree [24]	Standard parameters
Support Vector Machine [25]	kernel is linear
Gaussian Naïve Bayes [26]	Standard parameters
AdaBoost [27]	Number of estimators = 100
XGBoost [28]	Number of estimators = 100
Random Forest [29]	Number of estimators = 100

Table 3. Comparison of the Classification Algorithms Performance

Machine Learning Algorithms	Class	Precision	Recall	F1-Score	Accuracy
Logistic Regression	Non Poor	0.939037	0.940043	0.939540	0.909503
	Poor	0.821467	0.818856	0.820159	
Gradient Boosting Machine	Non Poor	0.972	0.972	0.972	0.972
	Poor	0.971774	0.971774	0.971774	
Stochastic Gradient Descent	Non Poor	0.853905	0.995004	0.919070	0.868927
	Poor	0.970894	0.494703	0.655439	
Passive Aggressive	Non Poor	0.937033	0.940043	0.938536	0.907902
	Poor	0.820321	0.812500	0.816392	
k-Nearest Neighbors	Non Poor	0.927124	0.930764	0.928940	0.893486
	Poor	0.792069	0.782839	0.787427	
Decision Tree	Non Poor	0.932056	0.920414	0.926199	0.890283
	Poor	0.772217	0.800847	0.786271	
Support Vector Machine	Non Poor	0.941778	0.952534	0.947126	0.920448
	Poor	0.854167	0.825212	0.839440	
Gaussian Naïve Bayes	Non Poor	0.914248	0.741970	0.819149	0.754939
	Poor	0.508832	0.793432	0.620033	
AdaBoost	Non Poor	0.946671	0.943969	0.945318	0.918313
	Poor	0.835084	0.842161	0.838608	
XGBoost	Non Poor	0.945083	0.939686	0.942377	0.914042
	Poor	0.823958	0.837924	0.830882	
Random Forest	Non Poor	0.893373	0.971806	0.930940	0.892152
	Poor	0.886819	0.655720	0.753959	

The first method, called random over-sampling, selects and duplicates a random subset of the minority class. In contrast, the second method is called random under-sampling and involves removing a random subset of the majority class. The third method is called Synthetic Minority Over-sampling Technique (SMOTE), and it works similarly to random over-sampling except that it creates a fraction of synthetic instances rather than replicating a random percentage of instances.

4.2.1. Random Oversampling Duplicating instances inside the minority class is how this approach operates. Therefore, a machine learning system will have a higher likelihood of detecting the minority class. A machine learning algorithm does better at spotting patterns that differentiate several classes when it oversamples. More significantly, no information is lost. Nevertheless, overfitting may become an issue if events within the minority class are repeated. Here, the two most effective algorithms that were previously presented—Gradient Boosting Machine and Support Vector Machine—are evaluated. A few performance metrics are evaluated after the random oversampling approach is used several times for varying numbers of instances. For the Gradient Boosting Machine and Support Vector Machine approaches, the optimal number of examples that yield the best F1-score is 13,000 and 17,000, respectively as shown in Figure 4.1.

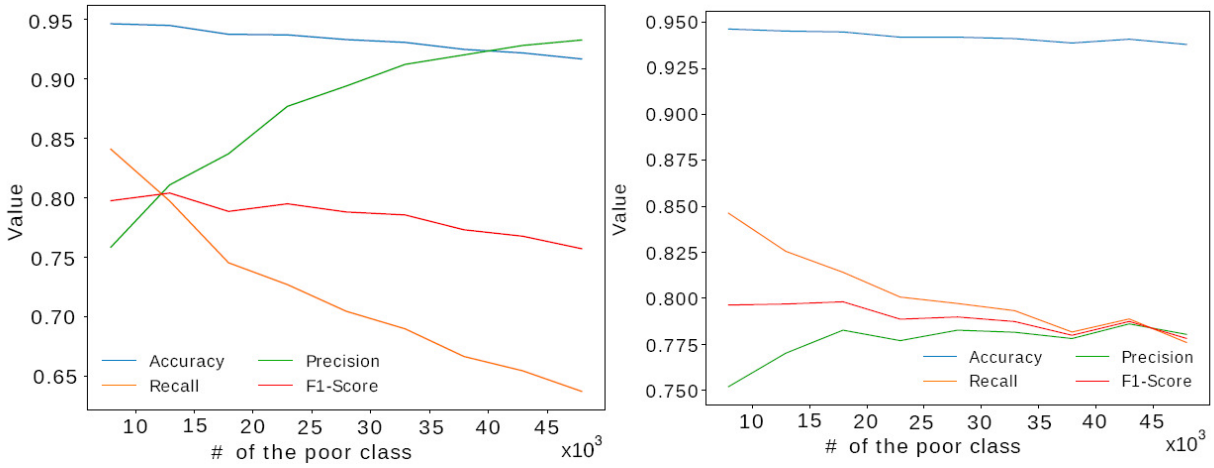

Figure 4.1. Evaluation of the Random Oversampling Procedure Performance

4.2.2. Random Undersampling To enhance a machine learning algorithm's performance, a random number of examples are removed from the majority class using this strategy. But this method has advantages and disadvantages of its own. Although there is a chance of information loss or, worse, the dataset becoming less representative, this strategy can be computationally efficient when the amount of training data is decreased. The random undersampling process is repeated and several performance metrics are evaluated, same as the oversampling technique. For the two algorithms—the Gradient Boosting Machine and the Support Vector Machine—the optimal F1-scores are produced by about 33,000 occurrences as shown in Figure 4.2.

4.2.3. Synthetic Minority Over-Sampling Technique (SMOTE) The true purpose of this method is to solve the difficult overfitting issue that may arise from using the random oversampling approach. This is accomplished by creating artificial instances rather than replicating pre-existing ones.

Nevertheless, because surrounding instances may belong to different classes, the SMOTE may inject noise into the data. The determination of a ratio parameter between 0 and 1 is one of the primary phases in the SMOTE. The algorithm is instructed to sample the minority class to a desired number of data points via this ratio parameter. However, utilizing SMOTE to increase the quantity of data points does not ensure correct findings. Therefore, it is necessary to experiment with various percentages. Grid search is a useful tool for this. In addition to using the f1-score as a performance assessment metric, a stratified 3-fold cross validation is employed. This approach uses the Gradient Boosting Machine because of its computational efficiency. It is evident that an optimal ratio parameter can be found around 0.19, which upsamples the minority class to about 10,340 (number of data points in the majority class 54,422 - optimal ratio parameter 0.19), despite the oscillatory F1-Score as the ratio parameter is increased from 0.18 to 1 as shown in Figure 4.3.

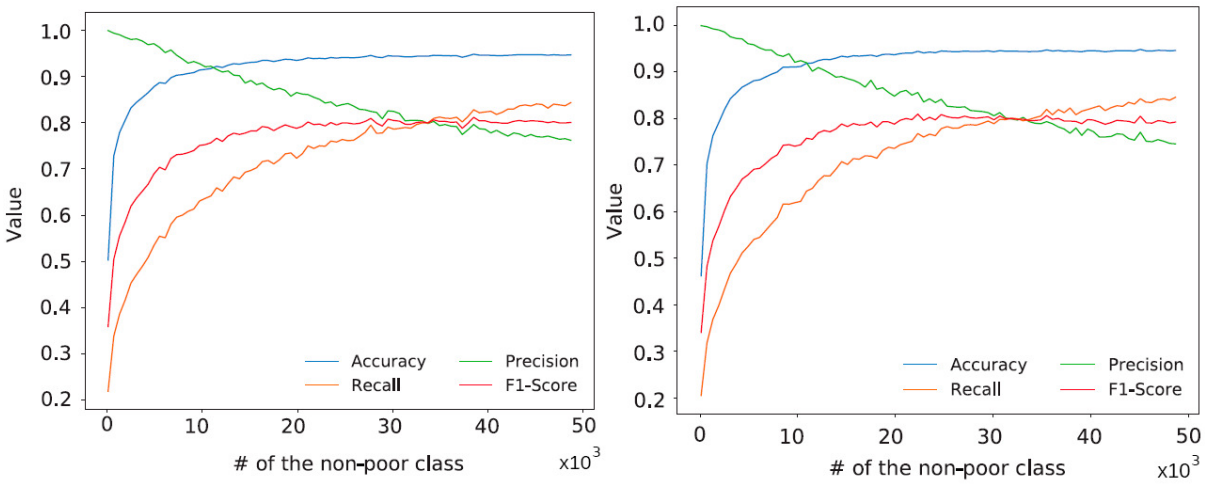

Figure 4.2. Evaluation of the Random Undersampling Procedure Performance

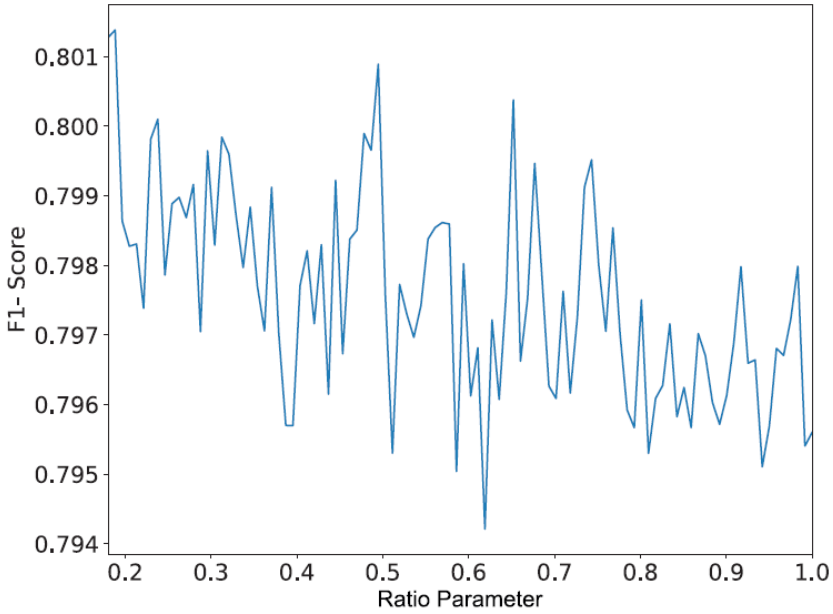

Figure 4.3. Performance of the Gradient Boosting Machine algorithm using SMOTE

5. Study Limitations

Below are some limitations that the authors of our study must deal with:

- The dataset is unbalanced, which can impact the performance of some machine learning algorithms.
- The study focuses only on Egypt and may not be generalizable to other countries.
- The authors acknowledge the need for further research to refine the models and address limitations.

6. Conclusion and future research

This research provides insight into Egypt's multi-dimensional poverty status of the Egyptian household. The highlights of this effort include analyzing and utilizing data on poverty from many national household surveys.

The use of machine learning techniques to evaluate and track Egyptian households' poverty levels is discussed. This method considers household expenditure and income surveys (HICES). The final model is simple to implement and use, allowing non-specialists to predict the likelihood that a household would be impoverished, the rate of poverty at any given time, and the rate of change over time. Various machine learning classification techniques are assessed. The Gradient Boosting Machine and support vector machine algorithms outperformed the other learning algorithms in terms of performance.

The datasest used is sufficient in its size and representative of all segments of Egyptian society. The data collected by the HICES. The dataset is prepared by converting categorical data into numerical form using one hot encoding in addition to bringing all the features on comparable scale. The latter is approached by standardization. Standardization resulted in a slightly better performance, and thus it has been adopted. The class imbalance problem that the dataset suffers from has been also investigated. Random oversampling, random undersampling and SMOTE were applied for this challenging task. Future research should take into account the needs for accuracy, training duration, and interpretability while selecting an algorithm. In this sense, finding the best prediction model is an iterative process that needs to be repeated with new data and new methods, on a regular basis. Possibilities for future research include, for instance, an application of deep learning models, other state-of-the-art algorithms. Also, it can be compared then with machine learning models.

Also, further research could explore ways to address the issue of unbalanced data, such as using oversampling or undersampling techniques. Additional studies could examine the performance of the models in other countries to assess their generalizability. Further refinement of the models, including feature selection and hyperparameter tuning, could improve their accuracy and efficiency.

REFERENCES

1. A. F. Psaros, X. Meng, Z. Zou, L. Guo, and G. E. Karniadakis, *Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons*, Journal of Computational Physics, vol. 477, pp. 111902, 2023.
2. H. Nozari, J. Ghahremani-Nahr, and A. Szmelter-Jarosz, *AI and machine learning for real-world problems*, In Advances In Computers, vol. 134, pp. 1–12, Elsevier, 2024.
3. S. Kochukrishnan, P. Krishnamurthy, and N. Kaliappan, *Comprehensive study on the Python-based regression machine learning models for prediction of uniaxial compressive strength using multiple parameters in Charnockite rocks*, Scientific Reports, vol. 14, no. 1, pp. 7360, Nature Publishing Group UK London, 2024.
4. B. B. Hazarika, D. Gupta, and P. Borah, *An intuitionistic fuzzy kernel ridge regression classifier for binary classification*, Applied Soft Computing, vol. 112, pp. 107816, Elsevier, 2021.
5. M. J. Abinash and V. Vasudevan, *A hybrid forward selection based lasso technique for liver cancer classification*, In Nanoelectronics, Circuits and Communication Systems: Proceeding of NCCS 2017, pp. 185–193, Springer, 2019.
6. M. Abualhaj, M. Al-Zyoud, M. Hiari, Y. Alrabanah, M. Anbar, A. Amer, and A. Al-Allawee, *A fine-tuning of decision tree classifier for ransomware detection based on memory data*, International Journal of Data and Network Science, vol. 8, no. 2, pp. 733–742, 2024.
7. S. Subbarayan, S. Thiyagarajan, S. Karuppannan, and B. Panneerselvam, *Enhancing groundwater vulnerability assessment: comparative study of three machine learning models and five classification schemes for Cuddalore district*, Environmental Research, vol. 242, pp. 117769, Elsevier, 2024.
8. Z. Khandezamin, M. Naderan, and M. J. Rashti, *Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier*, Journal of Biomedical Informatics, vol. 111, pp. 103591, Elsevier, 2020.
9. F. Kuran, G. Tanırcan, and E. Pashaei, *Performance evaluation of machine learning techniques in predicting cumulative absolute velocity*, Soil Dynamics and Earthquake Engineering, vol. 174, pp. 108175, Elsevier, 2023.
10. A. Alsharkawi, M. Al-Fetyani, M. Dawas, H. Saadeh, and M. Alyaman, *Poverty classification using machine learning: The case of Jordan*, Sustainability, vol. 13, no. 3, pp. 1412, MDPI, 2021.
11. N. Sani, M. A. Rahman, A. A. Bakar, S. Sahran, and H. M. Sarim, *Machine learning approach for bottom 40 percent households (B40) poverty classification*, Int. J. Adv. Sci. Eng. Inf. Technol, vol. 8, no. 4-2, pp. 1698, 2018.
12. Z. Huang, *Poverty Prediction Through Machine Learning*, In 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), pp. 314–324, IEEE, 2021.
13. M. T. Majeed and M. N. Malik, *Determinants of household poverty: Empirical evidence from Pakistan*, The Pakistan Development Review, pp. 701–717, 2015.

14. A. Sączewska-Piotrowska, *Determinants of the state of poverty using logistic regression*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, 2018.
15. X. Zheng, W. Zhang, H. Deng, and H. Zhang, *County-Level Poverty Evaluation Using Machine Learning, Nighttime Light, and Geospatial Data*, Remote Sensing, vol. 16, no. 6, pp. 962, MDPI, 2024.
16. L. Maruejols, H. Wang, Q. Zhao, Y. Bai, and L. Zhang, *Comparison of machine learning predictions of subjective poverty in rural China*, China Agricultural Economic Review, vol. 15, no. 2, pp. 379–399, Emerald Publishing Limited, 2023.
17. P. Corral Rodas, H. Henderson, and S. Segovia, *Poverty mapping in the age of machine learning*, Available at SSRN 4587156, 2023.
18. S. K. Satapathy, S. Saravanan, S. Mishra, and S. N. Mohanty, *A comparative analysis of multidimensional COVID-19 poverty determinants: An observational machine learning approach*, New generation computing, vol. 41, no. 1, pp. 155–184, Springer, 2023.
19. S. Chalichalamala, N. Govindan, and R. Kasarapu, *Logistic Regression Ensemble Classifier for Intrusion Detection System in Internet of Things*, Sensors, vol. 23, no. 23, pp. 9583, MDPI, 2023.
20. T. K. Dash, C. Chakraborty, S. Mahapatra, and G. Panda, *Gradient boosting machine and efficient combination of features for speech-based detection of COVID-19*, IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 11, pp. 5364–5371, IEEE, 2022.
21. N. Deepa, B. Prabadevi, P. K. Maddikunta, T. R. Gadekallu, T. Baker, M. A. Khan, and U. Tariq, *An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier*, The Journal of Supercomputing, vol. 77, pp. 1998–2017, Springer, 2021.
22. B. V. Kiranmayee, C. Suresh, and S. SreeRakshak, *Classification of the Suicide-Related Text Data Using Passive Aggressive Classifier*, In Sentimental Analysis and Deep Learning: Proceedings of ICSADL 2021, pp. 439–449, Springer, 2022.
23. Mohebbanaaz, L. V. Rajani Kumari, and Y. Padma Sai, *Classification of arrhythmia beats using optimized K-nearest neighbor classifier*, Computers, Materials Continua, vol. 68, no. 3, pp. 3299–3313, Tech Science Press, 2021.
24. B. Charbuty and A. Abdulazeez, *Classification based on decision tree algorithm for machine learning*, Journal of Applied Science and Technology Trends, vol. 2, no. 1, pp. 20–28, 2021.
25. X. Yan and H. Zhu, *A novel robust support vector machine classifier with feature mapping*, Knowledge-Based Systems, vol. 257, pp. 109928, Elsevier, 2022.
26. K. Wabang, O. D. Nurhayati, and others, *Application of the naive bayes classifier algorithm to classify community complaints*, Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 6, no. 5, pp. 872–876, 2022.
27. J.-K. Tsai and C.-H. Hung, *Improving AdaBoost classifier to predict enterprise performance after COVID-19*, Mathematics, vol. 9, no. 18, pp. 2215, MDPI, 2021.
28. C. N. Obiora, A. Ali, and A. N. Hasan, *Implementing extreme gradient boosting (xgboost) algorithm in predicting solar irradiance*, In 2021 IEEE PES/IAS PowerAfrica, pp. 1–5, IEEE, 2021.
29. V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, *AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes*, The Journal of Supercomputing, vol. 77, no. 5, pp. 5198–5219, Springer, 2021.
30. J. Ugirumurera, E. A. Bensen, J. Severino, and J. Sanyal, *Addressing bias in bagging and boosting regression models*, Scientific Reports, vol. 14, no. 1, pp. 18452, Nature Publishing Group UK London, 2024.
31. P. Myśliwiec, A. Kubit, and P. Szawara, *Optimization of 2024-T3 aluminum alloy friction stir welding using random forest, XGBoost, and MLP machine learning techniques*, Materials, vol. 17, no. 7, pp. 1452, MDPI, 2024.
32. Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, *An improved random forest based on the classification accuracy and correlation measurement of decision trees*, Expert Systems with Applications, vol. 237, pp. 121549, Elsevier, 2024.
33. D. S. Nugroho, I. F. Hanif, M. A. Hasbi, F. Fredianto, A. M. Saputra, and R. Zildjian, *Analisis Sentimen Dugaan Pelanggaran Pemilu 2024 Berdasarkan Tweet Menggunakan Algoritma Naïve Bayes Classifier: Sentiment Analysis of Alleged 2024 Election Fraud Based on Tweets Using the Naïve Bayes Classifier Algorithm*, MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 3, pp. 1169–1176, 2024.
34. M. Khashei, S. Etemadi, and N. Bakhtiarvand, *A New Discrete Learning-Based Logistic Regression Classifier for Bankruptcy Prediction*, Wireless Personal Communications, vol. 134, no. 2, pp. 1075–1092, Springer, 2024.
35. F. Nurrizky and S. Dwiasnati, *Comparison of Naive Bayes and Support Vector Machine (SVM) Algorithms Regarding The Popularity of Presidential Candidates In The Upcoming 2024 Presidential Election*, Computer Engineering and Applications Journal, vol. 13, no. 1, pp. 17–28, 2024.

Appendix:

Table 4. Features description

Name	Type	Description
Gender	Categorical	Gender of household head
Age	Numerical	Age of household head
Region	Categorical	Rural or urban
Work	Categorical	Does the household head work?
Smoke	Categorical	Does the household head smoke?
Rooms number	Numerical	Number of rooms in the house
hhsize	Numerical	Household size
Persons number	Numerical	Persons number per room
Household expenditure	Numerical	Household Expenditure per year
Person Expenditure	Numerical	Person Expenditure per year
Person Income	Numerical	Person income per year
Car	Categorical	Does the household own a car?
bicycle	Categorical	Does the household own a bicycle?
Motorcycle	Categorical	Does the household own a Motorcycle?
Telephone	Categorical	Does the household own a Telephone?
Mobile	Categorical	Does the household own a Mobile?
Smart Phone	Categorical	Does the household own a Smart Phone?
Internet	Categorical	Does the household own an Internet?
Refrigerator	Categorical	Does the household own a refrigerator?
Freezer	Categorical	Does the household own a Freezer?
Colder	Categorical	Does the household own a colder?
Oven	Categorical	Does the household own an oven?
Gas stove	Categorical	Does the household own a gas stove?
Microwave	Categorical	Does the household own a Microwave?
washing machine	Categorical	Does the household own a washing machine?
Semi-Automatic Washing Machine	Categorical	Does the household own a Semi-Automatic Washing Machine?
Full-Automatic Washing Machine	Categorical	Does the household own a full-Automatic Washing Machine?

Table 5. Features description

Name	Type	Description
Dishwasher	Categorical	Does the household own a Dishwasher?
Water_heater	Categorical	Does the household own a Water heater?
Vacuum_cleaner	Categorical	Does the household own a Vacuum cleaner?
Fan	Categorical	Does the household own a fan?
Heater	Categorical	Does the household own a heater?
electric iron	Categorical	Does the household own a electric iron?
LCD TV	Categorical	Does the household own an LCD TV?
black TV	Categorical	Does the household own a white black TV?
video DVD	Categorical	Does the household own a video DVD?
MP3 Player	Categorical	Does the household own an MP3 Player?
Casset	Categorical	Does the household own a Casset?
Desh	Categorical	Does the household own a Desh?
laptop	Categorical	Does the household own a laptop?
Tablet	Categorical	Does the household own a Tablet?
Camera	Categorical	Does the household own a camera?
Digital Camera	Categorical	Does the household own a digital Camera?
Filter of water	Categorical	Does the household own a filter of water?
Blender	Categorical	Does the household own a blender?
Kitchen machine	Categorical	Does the household own a kitchen machine?
Sewing machine	Categorical	Does the household own a sewing machine?
Electrical generator	Categorical	Does the household own an electrical generator?
Ration card	Categorical	Does the household own a ration card?
Fortified bread	Categorical	Does the household get fortified bread?
Education	Categorical	Does the household head expand on education?
Chronic illness	Categorical	Does the household head has long term or chronic illness?
Occasional illness or injurie	Categorical	Does the head household head has occasional illness or injuries?
Health	Numerical	The level of health difficulty
Poverity	Categorical	Is the household poor?