International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 9, Issue 1 (2023)
https://doi.org/10.31436/ijpcc.v9i1.326

The Forecasting of Poverty using the Ensemble Learning Classification Methods

Muhammad Haziq Adli Bin Zamzuri, Sofian Nadilah, Raini Hassan*
Department of Computer Science, Kulliyyah of Information and Communication Technology,
International Islamic University Malaysia, Kuala Lumpur, Malaysia

*Corresponding author: hrai@iium.edu.my
(Received: 29th May 2022; Accepted: 6th January 2023; Published on-line: 28th January 2023

Abstract— Poverty is a social-cultural problem that can be categorized into monetary approach, capability approach, social exclusion, and participatory poverty assessment. However, the existing measurement methods are complex, costly, and time-consuming. This research was conducted to forecast poverty using classification methods. Random Forest and Extreme Gradient Boosting (XGBoost) algorithms were applied to forecast poverty since they are supervised learning algorithms that use the ensemble learning approach for classification. Ensemble Learning has improved the classification of poverty and obtained better predictive performance. The results of the algorithms showed the poverty trend, which helped to determine the poverty classification. Hence, this method will help the government to act and produce a specific plan to reduce the poverty rate. It is a strategic move to reduce global poverty, parallel to Goal 1 of Sustainable Development Goal (SDG): No Poverty.

Keywords—Learning, Random Forest, Gradient Boosting, Extreme Gradient Boosting, XGBoost, Ensemble Learning Classification Methods, SDG.

I. INTRODUCTION

According to the 17 UN Sustainable Development Goals (SDG), poverty elimination is the first goal out of 17 goals [1]. Participating countries have contributed comprehensive efforts to manage the crisis and accomplish the goal of reducing the poverty rate while leaving no one behind which become a central promise from the 2030 agenda for Sustainable Development [2].

Since 2019, Covid-19 has left many countries economically shaken. [3] The global pandemic affected all sectors of economies across the globe and all layers of society which contribute to increasing poverty and unemployment. For example, business operations are forced to close during restricted movement, quarantine or lockdown which affects small businesses that depend on daily sales. This example shows that some groups of communities are more heavily affected by pandemics than others. The first goal of the UN Sustainable Development Goal; Ending Poverty by 2030 becomes a big new challenge as the global facing the crisis of economic fallout. It was stated that global poverty has been increasing for the first time since 1990 which is equivalent to a decade of effort to end poverty [4]. This factor increases the need to find the correct solution to reduce poverty. It is a vital issue as it is associated with other affairs such as lack of access to healthcare, security of the environment, homelessness, education, etc. All related issues will create dire consequences for every layer of society, economy, and country. This, increase the need for efforts to diminish the effects of the pandemic, especially for the needy.

To the best our knowledge, there are no direct or specific methods to measure the poverty level in every country. Traditionally, poverty measurement is based on an annual survey. This method is time-consuming and costly, which becomes a downside for the poorest or developing countries. Other than that, the data from the survey will only represent a small fraction of the households and it will be difficult to update the data regularly. In Malaysia, it was stated that the prediction task has not been implemented yet for the Bottom 10 Percent (B40) group [5]. By using the annual survey data, Malaysia has developed poverty line income (PLI) to classify the poverty group based on basic access such as health, education, living standard, etc. In addition, income is not the only factor that contributes to poverty. [6] There are groups of people that live above the poverty line but could not have basic access such as clean water, electricity, and education. These factors are considered non-monetary measurements. This highlights the importance of multidimensional poverty indicators.

In this research, the ensemble learning method will be used for forecasting poverty using the classification method. Using the accurate measurement method for poverty prediction is vital as the data can be used to solve the poverty issue. The data can be used to determine those who deserve to receive help from the government or to build a program that can be used to reduce the poverty rate in the country.

https://doi.org/10.31436/ijpcc.v9i1.326

II. Related Work

For this research, the researchers have studied numerous papers to learn methods and approaches that need to be taken to create proper research. Through the years, there are a lot of studies discussing poverty issues as it has become one of the major problems for many regions. According to 17 UN Sustainable Development Goals for the year 2030, the elimination of poverty worldwide ranked first [7]. There is no specific way that could describe the concept of poverty. As explained by economists, poverty is a social-cultural problem that can be described into various categories such as monetary approach, capability approach, social exclusion, and poverty participatory assessment [5].

Poverty is a large-scale issue that needs to be measured using the wealthiness scale. Current poverty measurement such as surveys is difficult to simplify into categorization. For example, In Malaysia, despite the decrease in poverty in the year 1970 from 49.3% to 2016 with 0.4%; there is a need for a comprehensive and accurate poverty measurement as the reduction level of poverty of the poor has not been shown. It is crucial to find methods that can classify poverty and establish strong accuracy. In addition, Proxy Means Test (PMT) is one of the most popular methods to measure poverty [8]. It was stated that the method is not accurate as most of the income and expenditure are understated. Therefore, an appropriate method of estimation is needed to measure the poverty level to help the target group in society.

Poverty classification and prediction are inconvenient for developing countries [6]. One of the reasons for its complications is data security and shortage which contribute to the inaccuracy of prediction or classification. For example, it was stated that classical methods such as surveys are not enough to help in decision-making by the government of Indonesia to classify households into categories [9]. Besides, geographical location and time are the aspects of poverty which is known as a heterogeneous problem. Thus, a novel method to collect data and implement classification is needed. Moreover, this study [6] stated that there are two types of complications of poverty measurement. The first measurement is poverty identification which has been solved traditionally by income. Meanwhile, the second complication is the creation of an index to measure poverty. The majority numbers of practitioners and researchers have used the multidimensional poverty index (MPI) to solve the second complication.

Studies regarding poverty prediction have been conducted using multiple methods and datasets. For example, in Malaysia, a statistical method has been used to classify poverty. Malaysia measured poverty using income indices or financial. Over 5 years, the Household expenditure, and Income Survey (HEIS) will be held twice. By using the data gained from the survey, the government has generated Poverty Line Income (PLI) and has listed multidimensional indicators to classify poverty. In a study, it was stated that by using machine learning, there is a possibility to use the data from eKasih to classify new poverty indicators. eKasih is a Poverty bank of Malaysia that was developed in 2017 to keep information about poverty in Malaysia [1]. Machine learning is a subset of artificial intelligence programs. [5] This study again stated that the program will be created to learn by themselves using algorithms that are done using the patterns of sets of data. As stated in the study by [7], it was explained that without explicitly instructing or programming the rule, machine learning will be able to train machines to learn and solve problems. Another example of the poverty measurement method is explained by researchers in a study [10], where it was stated that in recent years, mobile data and satellite images are frequently used as data sources for the estimation of poverty. For example, to predict socioeconomic levels in a major city in a Latin-American country, call data record datasets using 38 features were used by implemented Support vector machine (SVM) and random forest, which contribute over 80% of the accurate classification rate.

Next, there are two categories of machine learning. The first one is supervised learning. It is a technique to map functions between input and output variables using labelled data [7]. From the provided input and desired output, supervised learning will learn a function that could match up both categorized data and provide relationships from it. In contrast, unsupervised learning is a machine learning technique that analyses the data using unlabelled data. According to a study [11], the supervised techniques function in two modules training and testing. Data patterns will be evaluated during training progress. Meanwhile, a similar pattern of data will be recognized during the testing process. This study concluded that supervised learning could improve the performance and the accuracy of results, which contrasts with unsupervised learning. Supervised learning can be categorized into 2 types which are classification and regression. The classification method will be used when the desired category of output has been made. Hence, it is suitable to use as a method that can provide a prediction of poverty by its classification. There are various algorithms implemented to perform classification such as Naive Bayes Classifier (NBS), Logistic regression, Decision Tree, Artificial Neural Network, and random forest [5].

Research has provided studies regarding classification methods by comparing the performance of random forest and decision trees [1]. Data preparation has been made by using data from the Information Coordination Unit, Prime Minister Department (ICU JPM) known as eKasih year 2017.

https://doi.org/10.31436/ijpcc.v9i1.326

A total of 15 variables w selected. The measurement of performance has been done to measure accuracy, confusion matrix and receiver operating characteristic (ROC). For the confusion matrix, random forest generated an accuracy of 99% with a processing time of 31.64s while for the Decision Tree, 98% with 3.34s of processing time. The Area Under the Curve (AUC) is a measurement method used in ROC. AUC results are used to distinguish classes. Thus, the better result of performance needs to be higher in value. Using ROC shows that the forest model is better than decision trees with an AUC value of 0.9999 compared with 0.9975 by the Decision Tree. Furthermore, the study [11] shows that random forest outperformed other machine learning classification algorithms such as Support Vector Machine and Logistic Regression. To identify the poverty level for different blocks of groups in the United States, random forest is the fastest and the most accurate compared to others. Other than that, the advantage of random forest is its scalability which can be utilized for bigger datasets and many dimensions or features.

The research conducted in 2018 shows that a lot of studies are missing important concepts in machine learning [5]. For example, Parameter tuning, feature selection methods and feature engineering. By using eKasih dataset, data pre-processing has been made in this study such as data cleaning, feature engineering, normalization, and feature selection. Next, the chosen algorithm to implement classification were Naïve Bayes, Decision Tree, and k-Nearest Neighbors. Each classifier is tuned using a variety of techniques. In this study, Discretization, Confidence Factor, Minimum Number of Objects, k-Value, Distance Function and Regularization have been used for tuning parameters to provide high-accuracy results. The overall result from the study shows that by implementing the methods mentioned, there are significant improvements in terms of accuracy. Out of three algorithms, the Decision Tree outperformed other classifiers.

The paper conducted by Sudesh Kumar focused on prediction tasks [7]. The study aimed to build supervised inductive learning models that implement classification methods to forecast the poverty level of a household. Classification algorithms that have been chosen are Logistic Regression, random forest, and Light gradient-boosting methods (LightGBM). The steps taken for the research are data collection, data cleaning and pre-processing, feature extraction, splitting the data and performance metrics. The result from the study shows that LightGBM outperformed other classifiers. Other than that, it shows random forest as the next best alternative algorithm that can be used to measure poverty classification. The weakness of this research is the lack of data used as compared to the other related data science research. In another study [12] showed that random forest and gradient boosting outperformed other algorithms with higher accuracy. The study compared multiple algorithms such as Decision Trees and Neural Networks. In addition, another outcome provided from this study was gradient boosting showed the highest performance.

Random forest is a good algorithm for classification and prediction, which has been used for various predictive studies in multiple fields such as in medical, economics and finance by using one-year data, random forest generates high accuracy in both rural and urban areas, as compared to linear regression [13], [14]. The random forest also has a satisfactory performance at the national level. In addition, by limiting the variables which illustrate a realistic scenario; where poverty is commonly tracked using limited data and surveys, the random forest provides outcomes without reducing accuracy as compared to a full model without restriction. Thus, proving that random forest is suitable to be implemented as an alternative for prediction in a real setting such as using a small survey with limited data. Furthermore, based on the study they concluded that "the random forest algorithm tends to outperform other machine learning methods due to its capability to fit complex association structures even with small datasets." Besides, their experiment showed that the algorithm has the best prediction method for income poverty rate and multidimensional poverty as compared to other methods such as neural network, support vector machine and generalized least squares.

Based on all related research, we conclude that the existing method to measure poverty such as surveys is difficult as there are a lot of factors that they need to be considered. As stated in one of the related studies [1], to determine the poverty level, income is not the only measurement. There are a lot of factors need to be considered such as the measurement of a country's health and education. Government must collect various sorts of data such as income, occupation, demographic, health, and members from lots of households from various places and states. Thus, these factors raised many challenges. For instance, certain areas such as towns are approachable and can be reached using normal access of transportation while some places such as rural areas required more expense for the workforce, transportation, accommodation, etc.

On top of that, some households may also not provide accurate income and expenditure information as well [8]. Therefore, there are possibilities that the obtained data were not accurate thus government or policy maker could not solely rely on survey-based data for decision-making. Based on these factors, the researchers conclude that the existing method of poverty measurement can be improved by implementing poverty prediction using a machine learning approach. By using existing data sources, an

https://doi.org/10.31436/ijpcc.v9i1.326

accurate outcome of poverty prediction could be obtained. Overall, it can be said that this method can do a better job compared to the existing method since this method is scalable, faster, and cheaper.

In addition, the literature review also revealed that it is important to choose the right algorithm for the classification and forecasting method. Choosing the right method such as steps in data pre-processing is vital to produce high accuracy of prediction which has been explained in the study [5]. By using the accurate method, the measurement result can be used to develop an empowering policy. As a result, accurate prediction can be used to step up efforts on reducing poverty by identifying the group of communities that are in need. As explained by Sarwosri et al [9], the accurate outcome of poverty classification and prediction will be used by the local government to design a suitable program and help in decision-making to deliver targeted programs to targeted groups.

III. Methodology

In this research, a proper methodology has been implemented to gain accurate results in the prediction and classification of poverty. As shown in Figure 1, this method started with data collection and process with data pre-processing, Exploratory Data Analysis (EDA), data splitting, models training and testing and lastly, the model evaluation.

Python programming language was used in this research to do all the processes mentioned briefly above. Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. The researchers will be using the tool to perform necessary processes to the dataset before using it to train and test the models. There are some common packages that can work with Python such as Pandas, NumPy, Scikit-Learn, Matplotlib, Seaborn, etc. Figure 1 shows the methodology and overall process of this research.

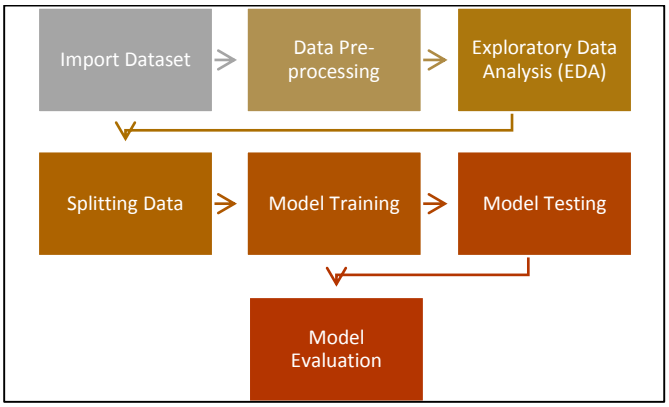

Fig 1: Flow of Process

A. *Dataset*

There are several datasets that have been used in our research. The first dataset is named "Poverty and Equity Database from World Bank Open Data" which was acquired from Kaggle which is a website of online datasets that stores many datasets. This website contains a collection of domain theories, databases, and data generators that are used and can be used by the community for further analysis. The dataset contains the latest poverty and inequality indicators compiled from officially recognized international sources.

Poverty indicators include the poverty headcount ratio, the poverty gap, and the number of poor at both international and national poverty lines. Inequality indicators include the Gini index and income or consumption distributions. The database includes national, regional, and global estimates. This database is maintained by the Global Poverty Working Group (GPWG), a team of poverty measurement experts from the Poverty Reduction and Equity Network, the Development Research Group, and the Development Data Group. This dataset was used to visualize world poverty from 1974 until 2015.

A survey was also conducted to gather some data that is not available online. In the survey, the respondents were asked to fill in some information on age, gender, ethnicity, state, strata, monthly income, and occupation. There is a total of 134 people who responded to this survey. After the data collection period was over, the respondent's information was exported to an excel file that contains 8 columns and 135 rows. This dataset was used to train a machine learning model to forecast poverty using ensemble learning classification methods which are Random Forest and XGBoost algorithms.

The other 3 datasets were from data.gov.my named "Incidence of Poverty by Ethnicity (2002-2019)", "Incidence of Poverty by State (2002-2019)" and "Incidence of Poverty by Strata (2002-2019)". These datasets contain data on the incidence of poverty in Malaysia from 2002 until 2019 which are separated into 3 parts: Ethnicity, State and Strata. These 3 datasets were used to visualize the actual poverty incidence in Malaysia from 2002 until 2019 based on states, strata, and ethnicity.

All datasets will be imported from csv/excel files to Panda's data frame in Python programming to further be processed later. Table 1 displays all the collected datasets and the variables used in the datasets with their description.

https://doi.org/10.31436/ijpcc.v9i1.326

TABLE I
COLLECTED DATASETS

Dataset	Feature Used	Description
Poverty and Equity Dataset	SP.POP.TOTL (Total Population)	Total world population from 1974 until 2015 (billions)
	SI.POV.NOP1 (Number of poor at $1.90 a day)	Number of poor people around the world at $1.90 a day from 1974 until 2015 (billions)
Survey Dataset	Gender	The survey respondent's gender (male/female)
	Age	The survey respondent's age
	Ethnic	The survey respondent's ethnic
	States	The survey respondent's state
	Strata	The survey respondent's strata (rural/urban)
	Total Monthly Income	The survey respondent's monthly income
	Occupation	The survey respondent's occupation
Incidence of Poverty by Ethnicity	Ethnic	List of ethnics in Malaysia
	Year	Year from 2002 until 2019
	Poverty Incidence	Incidence of poverty by ethnicity in Malaysia from 2002 until 2019
Incidence of Poverty by State	Country/State	List of states in Malaysia
	Year	Year from 2002 until 2019
	Poverty Incidence	Incidence of poverty by state in Malaysia from 2002 until 2019
Incidence of Poverty by Strata	Strata	Strata either urban or rural area
	Year	Year from 2002 until 2019
	Poverty Incidence	Incidence of poverty by strata in Malaysia from 2002 until 2019

B. *Data Pre-processing*

To begin the data pre-processing process, all the datasets used will be cleaned to make sure that the information presented during the data visualization process is accurate. Then, some of the data may need to be converted to computable form before using them to train the model.

Data Cleaning

Data cleaning is where the dataset is examined to observe whether it has any missing values, outliers, and duplications. The data then will go through a cleaning process to ease the visualizing and modeling works later without any problems such as an inaccurate representation of data or prediction.

In the dataset obtained from the survey, there are only 13 missing values out of 134 rows. All these missing values occur in the column named 'Monthly Income' where the values are not stated. Since there are only a few missing values in the dataset, the rows with the missing values were omitted from the dataset. Next, the unnecessary columns such as irrelevant columns that might not be important to the research will be discarded from the dataset. Lastly, the dataset used has a mixture of English and Malay Language. Hence, the variables in the Malay Language were translated into the English Language for convenience.

The other datasets such as "Poverty and Equity Database from World Bank Open Data," "Incidence of Poverty by Ethnicity (2002-2019)", "Incidence of Poverty by State (2002-2019)" and "Incidence of Poverty by Strata (2002-2019)", there are no missing values occurred in those datasets but some of the unnecessary columns were dropped from the datasets.

Data Transformation

Data transformation is the process where the data is converted from one format to another. This step is important because some machine learning algorithms cannot handle categorical variables for example Random Forest and XGBoost. Hence, data transformation is needed to encode them into numerical variables. This process can be achieved by using Label Encoder which is available in the Scikit-learn library. Label Encoder is a widespread practice to encode the string value in a categorical column into the machine-readable form that represents the category. After this process is done, the machine learning algorithms can fit and evaluate a model.

C. *Exploratory Data Analysis (EDA)*

Exploratory Data Analysis (EDA) is made to visualize the data to have a better understanding of the datasets and summarize their main characteristics. It also helps the researchers to gain some insights and statistical measures of the datasets.

Data Visualization

The Poverty and Equity Dataset from World Bank Open Data was used to illustrate world poverty over the year. Figure 2 shows the world's population and the number of poor under $1.90 a day in billions from 1974 until 2015. Based on the figure, even though the world's population is increasing from 1974 until 2015, the number of poor under $1.90 a day is showing a downward trend indicating that poverty is decreasing over the year.

https://doi.org/10.31436/ijpcc.v9i1.326

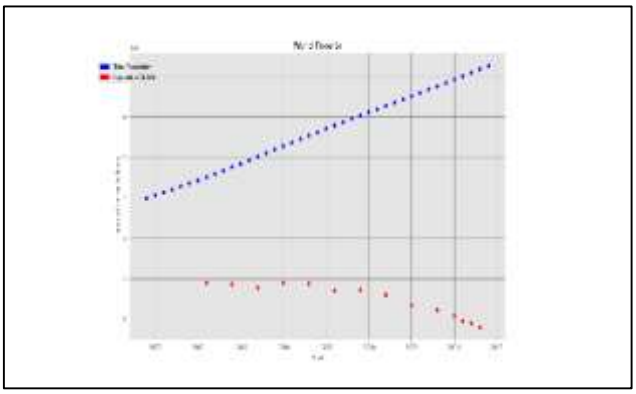

Fig.2: World Poverty

The Incidence of Poverty by State Dataset which was acquired from the Malaysian Open Data Portal was used to show the incidence of poverty in Malaysia. Figure 3 illustrates the poverty trend in Malaysia from 2002 until 2019.

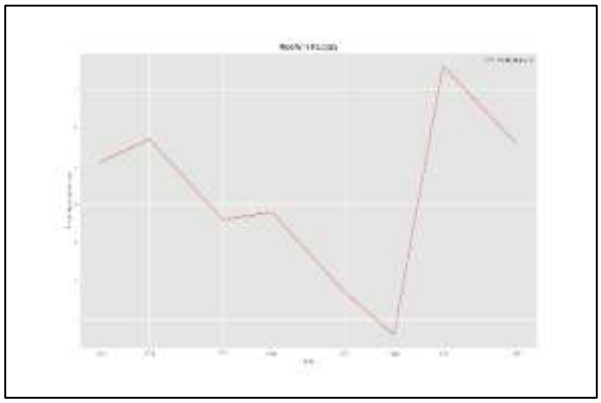

Fig.3 Poverty in Malaysia (Incidence of Poverty by State Dataset)

The previous dataset was also used to represent the poverty incidence in Malaysia according to the states. There are a total of 13 states including 3 Federal Territories which are Kuala Lumpur, Labuan, and Putrajaya. Figure 4 shows the poverty in Malaysia based on the state from 2002 until 2019.

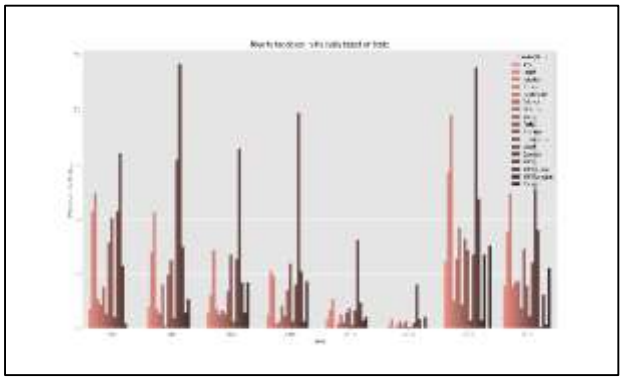

Fig.4 Poverty in Malaysia based on State (Incidence of Poverty by State)

Malaysian Open Data Portal also provided the Incidence of Poverty by Strata Dataset which was used to display the incidence of poverty in Malaysia according to urban and rural areas. Figure 5 demonstrates the poverty trend based on strata in Malaysia from 2002 until 2019.

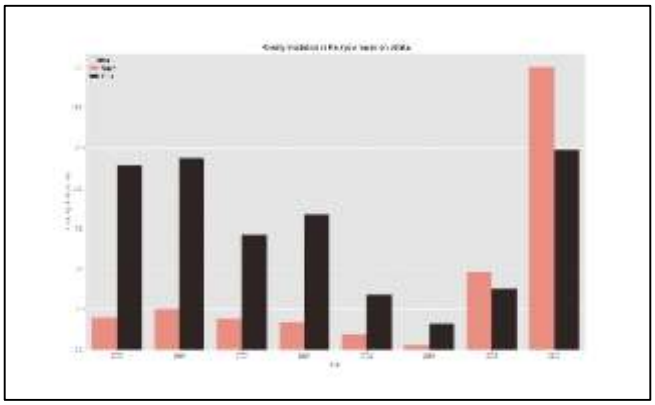

Fig.5 Poverty in Malaysia based on Strata (Incidence of Poverty by Strata Dataset)

The Incidence of Poverty by State Dataset which was also retrieved from the Malaysian Open Data Portal was used to exemplify the incidence of poverty in Malaysia based on ethnicity. There are 3 main ethnicities in Malaysia which are Bumiputera (Malays), Chinese and Indians. There are many other ethnic groups in Malaysia such as Iban, Kadazan, Melanau, etc. Figure 6 illustrates the poverty trend in Malaysia based on ethnicity from 2002 until 2019 (see Figure 6).

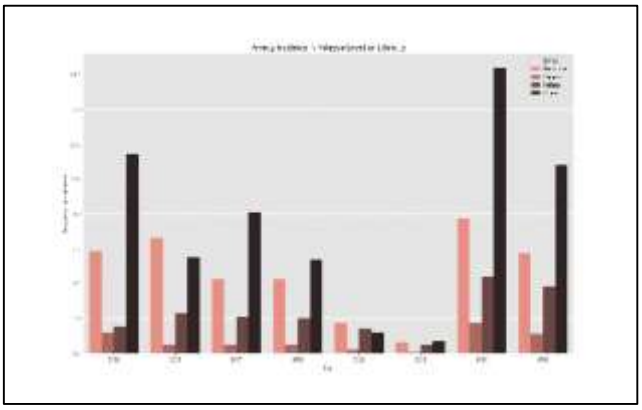

Fig.6 Poverty in Malaysia based on Ethnicity (Incidence of Poverty by Ethnicity Dataset)

The survey dataset which was obtained by conducting a survey was used to represent some poverty statistics in Malaysia. Figure 7 shows the poverty in Malaysia based on the survey dataset. From the figure, there are a total of 28 poor people from a total of 121 people.

https://doi.org/10.31436/ijpcc.v9i1.326

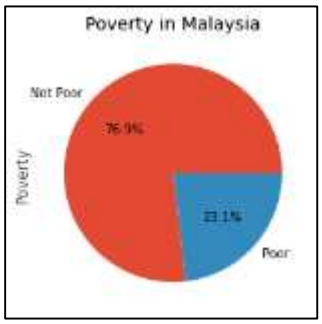

Fig.7 Poverty in Malaysia (Survey Dataset)

From the same dataset, the poverty incidence in rural and urban areas was demonstrated. Figure 8 represents the percentage of poverty in rural areas and urban areas separately. In rural areas, there are 9 poor people from a total of 35 people while in urban areas, there are 19 poor people from a total of 86 people.

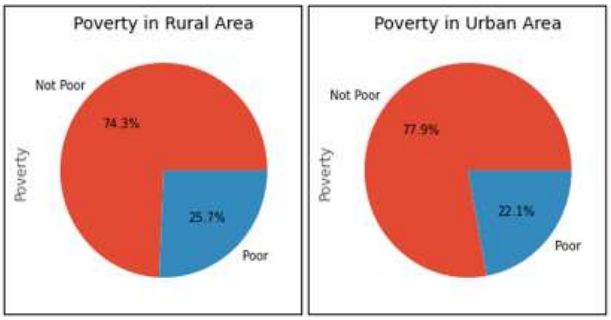

Fig.8 Poverty in Rural and Urban Area (Survey Dataset)

Lastly, the survey dataset was also used to show the poverty incidence between genders. Figure 9 displays the poverty cases between the male and female gender separately. Among male respondents, 17 of them are suffering from poverty with a total of 48 male respondents while among female respondents, only 11 of them are suffering from poverty with a total of 73 female respondents.

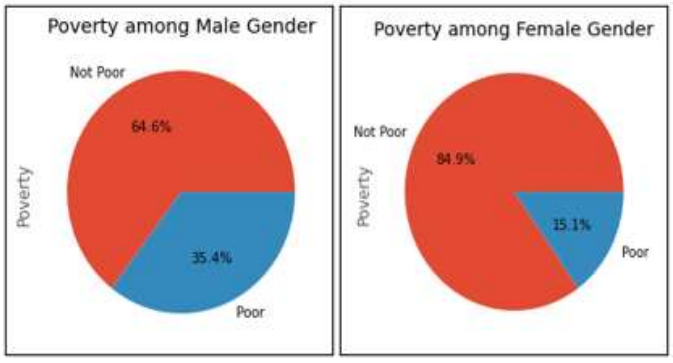

Fig.9 Poverty among Male and Female (Survey Dataset)

D. Feature Selection and Modelling

The feature selection process is conducted to find the correlation between the variables in the dataset. The highly correlated variables will be chosen to be included in the model training while the unnecessary variables will be excluded since they can lead the algorithm to yield garbage output in the end. This process is one of the crucial steps before proceeding to the model training process because too many variables can reduce computational performance. It also can cause high complexity that will extend the time to train the model.

Choosing a suitable algorithm to perform the feature selection technique is particularly important. The target variable and most of the predictor variables in the survey dataset are categorical variables. Thus, the researchers decided to apply the Chi-Square Feature Selection technique because this is one of the most suitable methods for datasets with categorical variables (see Figure 10).

The Chi-Square test is used to determine the correlation between the predictor and target variables. A low p-value indicates that the predictor variable has a high correlation to the target variable. Hence, the predictor variables with a low p-value will be selected to be trained in the machine learning model. The formula for the Chi-Square statistics used in the Chi-Square test is as indicated in (1).

$$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:
c = degree of freedom
O = observed value(s)
E = expected value(s)

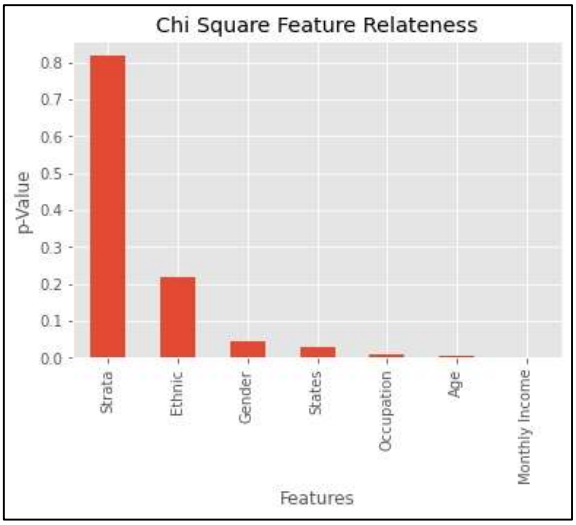

Fig.10 Chi-Square Test Result

For the model development, the research will be using two different machine learning models that will be compared in terms of their performance. The purpose of

https://doi.org/10.31436/ijpcc.v9i1.326

using two different models is to determine which model is suitable to train the dataset chosen. The models used are:

Random Forest

Random Forest is a supervised machine learning model that is widely known to have the capabilities in solving both classification and regression problems. Random Forest is one of the ensembles learning algorithms that create random decision trees from the sample data. Each tree then will predict the outcome and vote for the best solution in the end. Random Forest emphasizes the importance of features in determining the outcome of the model.

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting is also a supervised machine learning algorithm that have the same capabilities as the Random Forest algorithm which can solve regression and classification problems. XGBoost is another example of an ensemble learning algorithm that aggregates the ensemble of weak individual models to acquire a better final model. XGBoost implemented the gradient-boosted decision trees and was designed for a fast run time and high performance.

IV. Evaluation Results and Discussion

The performance of the model can be shown based on accuracy, recall, precision, and f1-score. Table 2: Summary of Results Table 2 displays the summary of results between two models used which are XGBoost and Random Forest.

Table IIIII
Summary of Results

Algorithm	F1-Score	Precision	Recall	Accuracy
XGBoost	1.00	1.00	1 .00	1.00
Random Forest	1.00	1.00	1.00	1.00

From Table 2, both XGBoost and Random Forest algorithms are showing the same performance with a perfect score for all evaluation metrics which are accuracy, recall, precision, and f1-score.

V. Discussion of Results

Based on the result of exploratory analysis, the percentage of poverty in rural areas is higher than the urban area. Next, the percentage of poverty incidence among male respondents is significantly more than the female percentage with a 20.3% difference.

For the feature selection part, by using the Chi-Square Test, it appears that the "Monthly Income" is the feature that has the highest correlation and most highly affects the target variable, "Poverty" with a p-value less than 0.001. The rank is then followed by "Age" with a p-value of 0.005, "Occupation" with a p-value of 0.008, "States" with a p-value of 0.028, "Gender" with a p-value of 0.044, and "Ethnic" with a p-value of 0.216. The last one in the rank is "Strata" with a p-value of 0.819 which is the feature with the lowest correlation to "Poverty" and can be considered as the least relevant to the target variable. From 7 available features, 4 of them were chosen to be the predictor variables to be trained into the model which are "Monthly Income", "Age", "Occupation", and "States".

Lastly, for the predictive analysis, before model training begins, the dataset was split into two parts which are train and test data. 70% of the data were used as training sets while the rest 30% were used as testing. From the sklearn library, two classifiers were imported to implement the model training which are Random Forest and XGBoost classifiers. As a result, both machine learning models achieved 100% accuracy along with 100% precision, recall and f1-score. The feature that affected these results is "Monthly Income" as the feature is highly correlated to the target variable with a p-value lower than 0.001 according to the Chi-Square test.

VI. Summary

Based on the result of this research, it can be deduced that random forest and XGBoost are applicable as classification models for poverty prediction. Regarding this research, it is shown that XGBoost and random forest algorithm produced the same result which is 100% accuracy. The reason behind the perfect accuracy is due to one of the predictor variables which is "Monthly Income" in the survey dataset had an incredibly significant relationship with the target variable which is "Poverty" with a p-value lower than 0.001 in the Chi-Square feature relatedness test. If the "Monthly Income" was dropped from predictor variables, the accuracy for both models was reduced to 64.86% on the Random Forest algorithm and 62.16% on the XGBoost algorithm.

Furthermore, the results may vary according to the dataset size used to create the machine learning model. In this research, the dataset used is small and it might be one of the reasons why the accuracy is perfect on both models. If the larger dataset is used, the result might be different according to how the researchers apply the feature selection, cross-validation technique and train the model. Thus, as a result, this research specified the alternative method to forecast poverty using the ensemble learning classification method.

VII. Recommendation and Future Works

There are some flaws in this research and many aspects can be improved further in the future. For example, the survey dataset used in this research can be considered small as there are only 134 rows of respondents that answered the survey. Hence, for future work, it is recommended to use a bigger dataset to see the true performance between the

predictive models used. Other than that, future researchers might want to consider using different predictive models from the current research and compare the performance between them. Lastly, using other evaluation techniques might help the researchers to get a better picture of how the model is performing.

ACKNOWLEDGEMENT

All praise to almighty Allah SWT, the most merciful, for the infinite blessings and opportunities that have been given to the authors. The authors would like to express gratitude to everybody who was involved in the entire process of completing this research. In addition, it was a great honor to receive all assistance and opportunity from all lecturers and staff of Kulliyyah of Information Technology, International Islamic University Malaysia (IIUM).

CONFLICT OF INTEREST

The authors declare that there is no conflict of Interest

REFERENCES

[1] A. A. Bakar, R. Hamdan, and N. S. Sani, "Ensemble learning for multidimensional poverty classification," Sains Malaysian, vol. 49, no. 2, pp. 447–459, 2020, doi: 10.17576/jsm-2020-4902-24.

[2] United Nations, "A UN framework for the immediate response to Table of Contents," United Nations, no. April, 2020.

[3] U. Nations, "Shared Responsibility, Global Solidarity: Responding To the Socio-Economic Impacts of Covid-19," United Nations, no. March, pp. 1–26, 2020, [Online]. Available: https://www.un.org/sites/un2.un.org/files/sg_report_socio-economic_impact_of_covid19.pdf.

[4] A. Sumner, C. Hoy, and E. Ortiz-Juarez, "Estimates of the impact of COVID-19 on global poverty," UNU WIDER Work. Pap., no. April, pp. 1–9, 2020, [Online]. Available: https://doi.org/10.35188/UNU-WIDER/2020/800-9.

[5] N. S. 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. 1698–1705, 2018, doi: 10.18517/ijaseit.8.4-2.6829.

[6] J. H. Mohamud and O. N. Gerek, "Poverty level characterization via feature selection and machine learning," 27th Signal Process. Commun. Appl. Conf. SIU 2019, pp. 6–9, 2019, doi: 10.1109/SIU.2019.8806548.

[7] S. K. Venkatramolla, "Machine Learning and Data Science for a Household-Specific Poverty Level Prediction Task," 2019.

[8] P. Kambuya, "Better model selection for poverty targeting through machine learning: A case study in Thailand," Thail. World Econ., vol. 38, no. 1, pp. 91–116, 2020.

[9] Sarwosri, D. Sunaryono, R. J. Akbar, and R. D. Setiyawan, "Poverty classification using Analytic Hierarchy Process and k-means clustering," Proc. 2016 Int. Conf. Inf. Commun. Technol. Syst. ICTS 2016, pp. 266–269, 2017, doi: 10.1109/ICTS.2016.7910310.

[10] D. R. Wijaya, N. L. P. S. P. Paramita, A. Uluwiyah, M. Rheza, A. Zahara, and D. R. Puspita, "Estimating city-level poverty rate based on e-commerce data with machine learning," Electron. Commer. Res., no. 0123456789, 2020, doi: 10.1007/s10660-020-09424-1.

[11] J. A. Talingdan, "Performance comparison of different classification algorithms for household poverty classification," Proc. - 2019 4th Int. Conf. Inf. Syst. Eng. ICISE 2019, pp. 11–15, 2019, doi: 10.1109/ICISE.2019.00010.

[12] H. Zixi, "Poverty Prediction through Machine Learning," Proc. - 2nd Int. Conf. E-Commerce Internet Technol. ECIT 2021, pp. 314–324, 2021, doi: 10.1109/ECIT52743.2021.00073.

[13] G. Cicceri, G. Inserra, and M. Limosani, "A machine learning approach to forecast economic recessions-an Italian case study," Mathematics, vol. 8, no. 2, pp. 1–20, 2020, doi: 10.3390/math8020241.

[14] P. Gogas and T. Papadimitriou, "Machine Learning in Economics and Finance," Comput. Econ., vol. 57, no. 1, pp. 1–4, 2021, doi: 10.1007/s10614-021-10094-w.