ISSN: 2723-9535

Available online at www.HighTechJournal.org

HighTech and Innovation Journal

Vol. 5, No. 2, June, 2024

Evaluating Household Consumption Patterns: Comparative Analysis Using Ordinary Least Squares and Random Forest Regression Models

En Lee [1], Thian Song Ong [1*], Yvonne Lee [2]

[1] *Faculty of Information Science and Technology, Multimedia University, 75450 Melaka, Malaysia.*

[2] *Faculty of Management, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.*

Received 24 January 2024; Revised 19 May 2024; Accepted 24 May 2024; Published 01 June 2024

Abstract

This research aims to decompose the contribution of socioeconomic factors towards household consumption expenditure using a regression approach, with log per capita expenditure as the dependent variable. Our study stands out as the first to utilise SHAP analysis and Machine Learning models to analyse household consumption expenditure. We select both OLS (linear) and Random Forest (nonlinear) models to compare how they estimate consumption expenditure differently. Both models explain about 85% of the variation in log per capita expenditure. The SHAP analysis reveals the nonlinear relationships inside the Random Forest model. Several insightful findings were suggested that can be integrated into current policy-making. The results are as follows: (1) Both models agree that income, household size, and educational level are major factors in the purchasing power of household heads. (2) The Random Forest model demonstrated a nonlinear contribution of age and household size towards log per capita expenditure, contrasting with previous studies that treated them as linear. (3) Household heads with a higher income and educational level tend to spend more. (4) Current policy should consider focusing on households with larger sizes and lower incomes, who tend to spend more despite earning less, primarily by assisting them with non-cash transfers and subsidies.

Keywords: Household Consumption; Machine Learning; Linear Regression; Random Forest; Shapley Value.

1. Introduction

For many years, poverty, as commonly measured by income, has been at the forefront of social and economic policy debates [1, 2]. Absolute poverty describes a situation where households or individuals are unable to meet minimum levels of standard of living in terms of income, food, health care, shelter, and other needs [3]. As a nation aspiring to achieve high-income status, the Malaysian government has introduced a series of initiatives to overcome the nation's poverty issues, from the New Economic Policy (NEP) in the 1970s to the present Twelfth Malaysia Plan. The earliest and most universally recognised method for measuring poverty is Poverty Line Income (PLI) [3]. It is a basic threshold to determine whether a household has adequate means for survival. By summing the two PLI indicators, food and non-food, households with a total income less than the combined PLI are considered to be in poverty, while households with a total income less than the food PLI are known as households in absolute poverty. In 2019, the Malaysian government adjusted the PLI threshold to RM2208, compared to only RM980 in 2005. However, although the 2019 PLI methodology

* Corresponding author: tsong@mmu.edu.my

http://dx.doi.org/10.28991/HIJ-2024-05-02-019

includes a household's food and non-food consumption in measuring poverty, it is just too narrow to reflect the complexity of households living in poverty [4]. A holistic poverty measurement methodology that covers different socio-economic indicators must be considered for economic policy planning, such as the multidimensional poverty index (MPI), which utilises three dimensions: health, education, and standard of living [5]. Rahman et al. (2021) [6] proposed an improved Malaysia MPI framework to enhance the poverty indicator's predictive powers with indicators for three dimensions, namely education, living standards, and employment. The indicators are literacy, education level, sanitation, housing, access to television services, and assets owned.

The current approaches used in Malaysia focus on using household income to define what constitutes a poor household in Malaysia. Simply relying on income data is neither sufficient nor accurate, especially with the increase in Malaysian household debt levels [7]. Income alone may not be enough to track poor households, especially the urban poor. Various factors must be considered in constructing an accurate household poverty classification method. One of the limitations of the income-based approach is the varying cost of living across states and regions. Besides that, income variations due to age and life stage can have an impact on the accuracy of poverty estimates. For instance, a retired couple with little or no income but substantial savings or assets may have a higher standard of living than a younger household with more income but less savings or assets. Furthermore, income data may fail to account for differences in the cost of living across strata, for example, the rural-urban gap. Recent trends show that consumption or expenditure patterns could be a good indicator for measuring poverty. Household consumption expenditure, which includes spending on necessities such as food, education, and health, can help infer the level of deprivation based on the type and quantity of consumption expenditure. Analysing the proportion of household consumption expenditure on basic needs such as food can tell whether the household faced deprivation or otherwise. Bhanoji Rao (1981) [8] mentioned that calculating the deprivation point from annual expenditure can measure incidences of deprivation and thus construct the poverty line. Besides that, Kumar et al. (2009) [9] investigated the deprivation of food in India by looking at expenditures on cereal to understand whether the trend was declining or increasing before and during the India Reform Period to determine household poverty status.

Recently, empirical literature has begun to employ machine learning for analysing the factors that influence household income and expenditure [10–13]. Herrera et al. (2023) [10] utilised the traditional linear regression model as well as other machine learning regression models such as Elastic Net, XGBoost, and Neural Network to investigate the correlation between household socioeconomic and ICT characteristics and household income. It is important to mention that although Elastic Net is part of the machine learning field, it is a linear model. The findings indicate a correlation between educational attainment and the use of information and communication technologies (ICTs). Higher educational levels and an older average age of household members are associated with higher incomes. Among the 4 models, XGBoost and the neural network outperforms the rest in terms of accuracy. The authors emphasised the superiority of the two nonlinear machine learning models due to their ability to uncover nonlinear relationships between variables, which are often masked by the linear-based regression model. This can be verified by the SHAP summary plot inside the study, as the XGBoost and Neural Network models treated the education level (the top-ranked variable in all the proposed models) as an exponential function, while the linear and Elastic Net models assumed a linear relationship between the educational level and income. Overall, this study shows that individuals with higher levels of education tend to perform a broader range of complex tasks online and are more likely to gain more income. The finding further emphasises the role of education in driving socioeconomic outcomes.

Hwang et al. (2022) [11] proposed deep learning clustering and logistic regression models to analyse the heterogeneity in about 50,000 households in Korea. A clustering model was first constructed to examine the financial heterogeneity of households in 8 clusters. The primary factor for a household to fall into wealthy clusters is their assets in real estate, followed by loans obtained for real-estate investments. Later, demographical analysis was done by building a logistic regression model with household demographic variables (age, educational level, income level, and household size). Generally speaking, across the clusters, household heads with a higher age, better education, and higher income live in the wealthy clusters, and vice versa. Moreover, the authors also investigated the probability of certain households climbing from poor clusters to wealthy clusters from year 2017-2020. Unfortunately, those living in poor clusters are more likely to move within the four poor clusters only.

Chowdhury et al. (2023) [12] investigated the impact of BRAC's Ultra-Poor Graduation (UPG) model on the participant's wealth and expenditure level using Honest Causal Forest (HCF), one of the recent tree-based machine learning algorithms. The UPG model offered participants both consumption expenditure supports or a grant of productive assets with technical skills training. Findings show that the UPG programme led to significant gains by participants in either wealth accumulation or consumption gain only. The affected households with a higher gain in asset outcome are generally older, more dependent on wage income, and had less self-employment income at the baseline, while participants experiencing consumption gains that led to increased expenditures are younger and earned higher income from self-employment activities.

Zeng & Chen (2022) [13] studied the urban-rural integration types in China and their changes within a ten-year period from 1990 to 2020 using partitioning around medoids (PAM), a clustering-based machine learning model. Clustering was first done, and the authors concluded that the rural-urban transition should be represented in 4 clusters (high-level urban-rural integration, urban-rural integration in transition, early urban-rural integration in the backward stage, and low-level urban-rural integration). Overall, results suggest that urbanisation lifts up economic growth. The most important finding can be seen in Cluster 1. It is the highest urbanisation rate cluster, and the income, expenditure, housing areas, and educational level gaps are the lowest among all, implying negative relationships (the higher the urbanisation, the lower the gaps).

Although the importance of using consumption expenditure or income data for poverty analysis is recognised globally, research focusing on measuring poverty through consumption expenditure, especially in Malaysia, is still limited. To the best of the authors knowledge, the recent literature works that discuss household consumption issues with a regression approach in Malaysia are as follows: Ang & Cheah (2023), Zin & Nabilah (2015), and Ayyash & Sek (2020) [14–16]. Ang & Cheah (2023) [14] discussed the consumption inequality issue among different income groups in terms of consumption of pharmaceutical goods only. Although the study highlighted disparities in expenditure in this industry, the broader scope of the consequences of household sociodemographic characteristics on other expenditure types remains unknown. On the other hand, Zin & Nabilah (2015) [15] conducted linear and quantile regression to identify the factors that contribute to household expenditure across urban and rural areas, but only considered three quantiles (three expenditure levels). Moreover, the authors only ranked the determinants without showing their coefficients; the degree of the determinants' impacts remains unknown. Ayyash & Sek (2020) [16] proposed Fields' regression approach to decompose consumption inequality based on household demographical variables. However, the proposed regression model assumes linearity, which might not hold for certain important variables such as the age of the household head. Given this research constraint, the purpose of this study is to investigate the regression analysis of consumption patterns using a machine learning approach. We suggest using Random Forest because of its broad application. Random Forest, along with many other Machine Learning models, has the ability to handle data that exhibits nonlinearity. In fact, our finding shows that age has a nonlinear relationship with log per capita expenditure, contrasting with that which was identified as having a positive linear relationship [15, 16]. Meanwhile, the contribution of household size towards log per capita expenditure was also found to be complex, although closely to negatively linear. These nonlinearities are discussed in detail in Section 3. One may question the black box nature of the Random Forest model, with concerns about the lack of transparency that makes it challenging to 'view' the relationship among variables. To address this issue, we proposed another interpretability tool called SHapley Additive exPlanations (SHAP), which allows us to better understand and explain the complex relationships between variables within the Random Forest model.

In this work, we mainly compare our result with Ayyash & Sek's (2020) [16] study. This study can be viewed as an extension of Ayyash & Sek's (2020) [16] study, as both studies use the same data source (Household Expenditure Survey, HES), select the same target variable (per capita expenditure), and decompose the degree to which determinants contribute to household expenditure, but this study further explores the scope using Machine Learning model. HES is an official survey programme launched by the Malaysian Department of Statistics (DOSM) that is conducted twice every five years to collect individual information and expenditure patterns via personal interviews. Ayyash & Sek's (2020) [16] study was conducted using the 2014 version of the survey, while this study uses the 2019 version, which updates the previous study's findings in addition to introducing an expanded analytical methodology.

Therefore, this study would like to investigate and analyse the consumption expenditure pattern in Malaysia, with the following objectives: 1) To implement and evaluate the performance of Linear and Random Forest regression models. 2) To understand the determinants' importance and the relationships that exist between a set of determinants and household expenditure. 3) To compare the two models based on their respective findings.

The main contribution of this study can be summarised as follows: (1) To the best of the authors' knowledge, this is the first study to attempt to use Machine Learning model in conjunction with an explainable model to quantify and visualise the impact of household demographic factors on household expenditure in Malaysia. (2) Income is the most significant determinant in both models. Similar to Ayyash & Sek's (2020) [16] result, from our proposed Ordinary Least Squares (OLS) model, household size and educational level are the next two most influential factors that determine household expenditure, followed by ethnicity and regional variables. However, findings from Machine Learning model surprisingly indicate that regional variables contribute to household expenditure more than educational level, while the importance of household size remains unchanged. (3) Based on the SHAP results, variables such as age and household size are found to have nonlinear relationships towards per capita log expenditure, compared to positive linear relationships found by Zin and Nabilah (2015) as well as Ayyash & Sek (2020) [15, 16]. This suggests that when decomposing household expenditure using regression, one should consider a regression model that can handle the variables that appear to be nonlinear in nature. (4) Instead of interpreting the contribution of each determinant alone, SHAP allows interpretation in three dimensions, providing a detailed picture of the overall relationship. Moreover, conveying the results visually enhances interpretation for broader audiences, such as policymakers.

The remaining parts are organised as follows: Section 2 introduces the dataset and the methodology used; Section 3 discusses the results of the study; and Section 4 summarises the study.

2. Material and Methods

2.1. Overview

In general, there are several phases in this work: data preparation, feature engineering, modelling, and the evaluation phase. At the end, SHAP analysis will be performed to further explain how the chosen Machine Learning model influences its decision to make such predictions. During the data preparation phase, samples from household and member records will go through a series of pre-processing steps before being fed into prediction models. In the feature engineering phase, per capita income and expenditure are computed and converted into natural logarithm form. Next, during the modelling phase, two models will be selected, which are OLS and Random Forest. Later, in the evaluation phase, several metrics are chosen to compare and assess their performance. Figure 1 exhibits the flowchart of the methodology and outlines the specific processes involved in the process, starting with data collection and concluding with model evaluation and SHAP analysis.

2.2. Datasets

This study makes use of two dataset records: household and member datasets. Both datasets can work independently or be linked with a unique key index column called HID (household ID) present in each dataset. HID serves as a unique identifier for every household. The information about the two datasets utilised is explained in the following section.

a) Household Dataset

The first dataset is the household dataset, which has a total of 16,354 observations, each of which represents a single household. This dataset contains ten columns. The first column, HID, represents the unique household ID that uniquely identifies each row. The variables, data type, data format, and description are shown in Table 1.

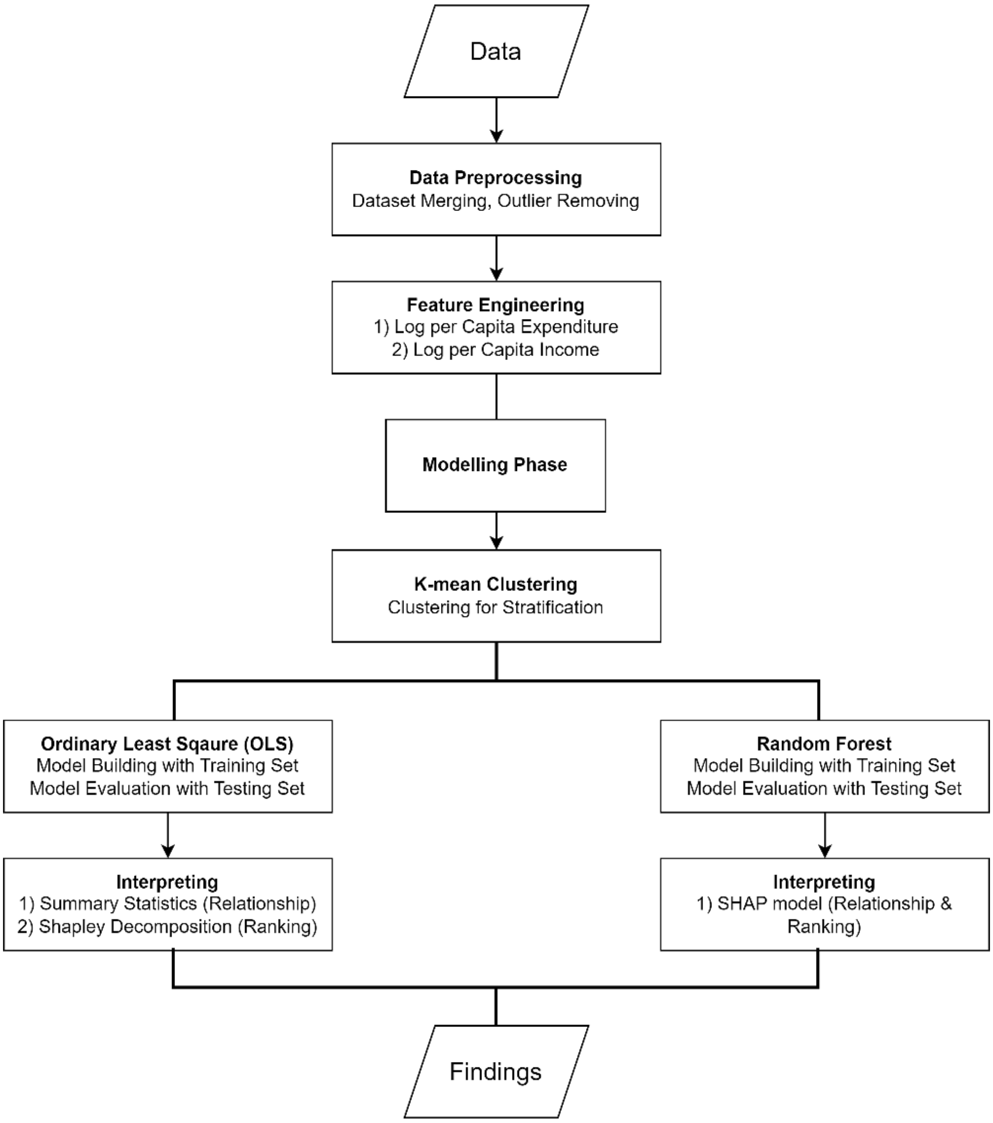
Data
Data Preprocessing
Dataset Merging, Outlier Removing
Feature Engineering
1) Log per Capita Expenditure
2) Log per Capita Income
Modelling Phase
K-mean Clustering
Clustering for Stratification
Ordinary Least Sqaure (OLS)
Model Building with Training Set
Model Evaluation with Testing Set
Random Forest
Model Building with Training Set
Model Evaluation with Testing Set
Interpreting
1) Summary Statistics (Relationship)
2) Shapley Decomposition (Ranking)
Interpreting
1) SHAP model (Relationship & Ranking)
Findings

Figure 1. Flowchart of the research methodology

Table 1. Variable in household dataset

Variable	Data Type	Data Format	Description
HID	object	28 Numerical objects	Unique Household ID
Weight	float	Numerical	Statistical adjustments that are made to household survey data
NO_HH	int	1-21	Household No
Saiz_HH	int	1-5	Total household members
State	int	1-16	Each represents a state/territory
Region	int	1-3	1= Peninsular Malaysia or 2= Sabah & Labuan 3 = Sarawak
Strata	int	1-2	1 = Urban 2 = rural
Ethnic	int	1-4	1 = Bumiputera 2 = Chinese 3 = Indian 4 = Others
Total_Exp_01_12	float	Numerical	Total Household Expenditure for 12 expense types
Total_Inc	float	Numerical	Monthly household gross income

b) Member Dataset

This dataset contains 214,719 observations, with samples gathered at the individual level. Because each person has a household ID, one or more people can originate from the same household. The variable, HID, which is the same as in the household dataset, serves as the key column to merge individual-level data with the household-level dataset. Among all observations, 64,160 individuals came from 16,354 different households. As a result, using the aggregation method, these 64,160 people can be combined into 16,354 households. The remaining 150,559 individuals, who represent 38,147 households, have no household-based data. Table 2 lists the dataset's variables, data type, data format, and description.

Table 2. Variable in member dataset

Variable	Data Type	Data Format	Description
HID	object	28 Numerical objects	Unique Household ID
NO_AIR	int	Numerical	Household Member No.
Relationship	int	1-12	Position of member in household
Sex	int	1-2	1 = Male 2 = Female
Age	int	0-98	00 = children < 1 year 01 = 1 to 97 years 98 = aged >= 98 years
Ethnic	int	1-4	1 = Bumiputera 2 = Chinese 3 = Indian 4 = Others
Marital_Status	int	1-5	1 = Never Married 2 = Married 3 = Widow/Widower 4 = Divorced 5 = Separated
Highest_Certificate	float	1-6	1 = No Certificate 2 = PMR/SRP 3 = SPM/ SPMV 4 = STPM 5 = Diploma / certificate 6 = Degree/Advance Diploma

Act_Status	int	1-15	1 = Employer 2 = Government employee 3 = Private employee 4 = Own account worker 5 = Unpaid family worker 6 = Unemployed 7 =Housewife/looking after home 8 = Student 9 = Government pensioner 10 = Private pensioner 11 = Elderly 12 = Persons with Disabilities 13 = Child not at school 14 = Infant 15 = Others
Income_Recipient	int	1-2	1 = Yes 2 = No
Occupation	int	1-3	01 = Manager 02 = Professional 03 = Technician and associate professionals

2.3. Data Preparation

The household and member records are made up of several household and member samples with various attributes that define each household's economic and non-economic condition. First, household and member records are merged into a dataset by using the variable HID as the key index. This joined dataset is based on the household dataset, which means it returns all household rows from the household table and matching records having the same households' ID from the member table. The remaining 38,147 households that have no household information are discarded. The merged dataset is a dataset with 64,160 rows and a total of 16,354 households inside. All records from the household dataset are retained, and then those records are matched with the same key index HID in the household dataset from the member dataset. Finally, only household head rows inside the table are retained, producing a dataset with 16,354 rows.

2.4. Modelling Phase

In this phase, two models, OLS and Random Forest, are chosen. Both models can be used to examine the relationship between household expenditure and its influencing factors. The Ordinary Least Squares represents a linear approach to regression, while the Random Forest offers nonlinear modelling that can capture nonlinear and complex relationships between predictors and the dependent variable. The aim of this work is to compare the performance of Random and OLS models, focusing on the models' interpretability of the nonlinearity connections and their overall accuracy scores.

c) Econometric Model

Model (1) is an application of OLS, which is a common econometric model used in economics. Following is the linear model built for this study:

$$log(y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + \in_i \tag{1}$$

where y_i denotes the dependent variable (per capita expenditure), $\log(y_i)$ is the natural logarithm form, X_i are independent variables, β_i are coefficients, β_0 is the intercept and $\in_i$ is the error term.

d) Random Forest Regression

This is one of the most common supervised Machine Learning algorithms that relies on ensemble learning to perform regression tasks. Multiple decision tree models predict the outcomes independently and then average them. For each decision tree model in the Random Forest model, subset of sample is selected independently to train it. Generally, Random Forest eliminates the overfitting problem due to its averaging properties. Random Forest can rank the feature importance by finding out their impurity decrease. The feature with the highest impurity decrease is the most significant feature. The equation of mean decreases the impurity measure as follows:

$$imp(X_m) = \frac{1}{N_T} \sum_T \sum_{t \epsilon T: \upsilon(s_t)=X_m} p(t) \Delta i(s_t, t) \tag{2}$$

where X_m represents a variable, $p(t)$ is the proportion of N_T/N of samples reaching t and $\upsilon(s_t)$ is the variable that was used when making a split, and $p(t)\Delta \mathrm{i}(s_t, t)$ is the total weighted impurity decrease for all nodes t when considering the X_m variable.

2.5. Experiment Setup

The dataset is initially divided into a training set and a testing set in a 4:1 ratio. This means that 80% of the data is used to train the model, with the remaining 20% used to test it. The performance of trained model will be evaluated based on the coefficient of determination or more commonly, R-squared (R^2). It measures the total explainable variance of predicted dependent variable from determinants. The formula of R^2 is as follows:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3}$$

where SS_{res} is the sum of the residuals squared, while SS_{tot} is the sum of distance the sample observations are from mean squared.

Besides that, mean squared error (MSE) also used to evaluate the model performance. It computes the average squared difference between the estimated value, $\hat{y}$ and the true value, y. The MSE can have only positive value, the closer the value to 0, the better the model performance. The MSE is calculated as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{4}$$

where y_i is true value, and $\hat{y}_i$ is the predicted value.

2.6. K-Means Clustering Approach for Stratification

The stratification process is done to ensure the training and testing sets have an equivalent proportion of expenditure groups. There is no such expenditure group inside the dataset; hence, clustering groups are created before performing stratification. K-means clustering is chosen to discover the clusters from these households that best represent the distribution. The K-means algorithm is an unsupervised learning method that iteratively assigns each of the observations to one of the clusters. The iteration stops when no further changes are found. The goal of K-means is to minimize the sum of Sum of Squared Error (SSE) between the data points inside the clusters. The formula is shown below:

$$SSE = \sum_{i=1}^{k}\sum_{o\in G_i} dist(o, cen_i)^2 \tag{5}$$

where k denoted the number of clusters, $o_1, \dots, o_n$ denoted the data, $G_1, \dots, G_k$ is the list of clusters, $cen_1, \dots, cen_k$ is the list of centroids from each cluster.

If there is no fixed number of k (cluster), the elbow method can help to determine the ideal k value. It works by looping through different values of k, then compute and plot the average SSE for corresponding k. The best k value is found at the elbow of the plot, and the decreasing effect of averaging SSE afterward is minor.

Silhouette is another way to find out the optimal k. In this method, silhouette score is computed to measure how closer an observation is within-cluster (cohesion) as compared to the neighbour clusters (separation). Silhouette score has a range of [-1 to 1], the closer the value to 1, the tightly the observation is to the centroid, 0 means that the observation is on a boundary that could be assigned to any two neighbouring clusters. -1 means observation assigned to a wrong cluster. Formula below shows the silhouette score, s for a single observation, i.

$$s(i) = \frac{b(i) - a(i)}{max\,\{a(i), b(i)\}} \tag{6}$$

where $a(i)$ is the average distance between observation i and all the other observations within the same cluster, $b(i)$ is the average distance from observation i to all the other clusters. The average silhouette score is then taken from all observations and repeated for all value of k to determine which number of k produce the highest score.

Though the primary goal of this research is to predict poverty in terms of consumption expenditure, analysing the results from k-means clustering will complement our findings by explaining the strength of determinants through a supervised regression approach. To achieve this, descriptive statistics are carried out to summarise the data from clusters, assess the similarity of the variables within the clusters, and identify any obvious differences between clusters.

2.7. Shapley and Owen Values

Shapley value is a game theory introduced by Lloyd Shapley. The idea is to fairly distribute the total gain to players of a game from the total contribution by them. In the regression field, R^2 measure the overall goodness of fit. However, being able to decompose the overall R^2 into individual R_i^2's represented by a single determinant is also desirable. To achieve this, the Shapley value is needed. It is represented by the partial, R_i^2 which is contributed by determinant x_i , and is given by following formula:

$$R_i^2 = \sum_{T\subseteq\{x_1,\dots,x_p\}\setminus\{x_i\}} \frac{k!(p-1-k)!}{p!}\left[R^2(T \cup \{x_i\}) - R^2(T)\right] \tag{7}$$

where T is a model trained with k determinants but without determinant x_i. $T \cup \{x_i\}$ representing model with all determinants (including x_i). p represent the number of determinants and k represent the subset of determinants used.

The Shapley value will be used to assess the determinants' contribution toward household expenditure in term of R_i^2. Summing up all the Shapley value is equivalent to the total explained variance, R^2.

The dummy variables will be examined further using the Owen value, which is an extension of the Shapley value. The method considers a set of determinants as a coalition structure and calculate the coalitional value of those determinants. Its concept is closely aligned with Shapley value. For the linear model's determinants, Shapley value will aid in estimating the partial R^2 for the linear model's determinants and ranking them based on their significant and contribution to the model.

2.8. SHAP

SHAP is an approach to explain the model's predictions by interpreting the features' contribution based on Shapley value. Lundberg & Lee (2017) [17] published the method, which has helped a lot of researchers discover the black box properties behind machine learning models. The benefits of SHAP are its global interpretability over the machine learning models through various plot analysis. The global explanation (which refers to several samples) is usually plotted with a bee swarm plot, commonly known as a summary plot in SHAP. Every dot inside the summary plot represents a Shapley value associated with a feature, and the colour represents the feature value. Besides that, the SHAP summary plot will rank the features according to their contribution. The SHAP dependence plot, a scatter plot that shows how various features influence the model's predictions, is another visualisation tool in our study. In this work, we employ SHAP to investigate the relationships between determinants and dependent variables, using both the dependence plot and summary plot for detailed insights.

3. Result and Discussion

3.1. Overview

In this section, we present our results and analysis from Sections 3.2 to 3.8. Subsequently, in Section 3.9, we discuss the findings by comparing between the OLS and Random Forest models used in this study. Additionally, we contrast these results with those from other studies. Following this comparison, we discuss the implications of our findings for existing literature and policy-making.

3.2. Pre-Processing for Modelling

As explained in Section 1, the dependent variable is household expenditure (*Total_Exp_01_12*) instead of income, due to a better reflection of the overall living standard of households. It is worth mentioning here that household expenditure is recorded for discretionary items and services, yet it does not encompass any investment allocations. The independent variables considered in this study are household size (*Saiz_HH*), educational level (*Highest_Certificate*), household income (*Total_Inc*), strata (*Strata*), ethnicity (*Ethnic*), region (*Region*). Total income is the monthly gross household income.

The dummy variables are created for ethnicity to facilitate regression analysis. The educational level is a ranking variable, which consists of six categorical values. Household size consists of five categorical values, ranging from 1 to 5. A household size of 5 also captures households with more than 5 members. Outliers of household expenditure are removed by removing the top 5% of the household expenditure distribution. Meanwhile, those households with household expenditures greater than their income were discarded. Following these changes, the dataset now contains 14,525 households, indicating a reduction of 1,829 households from the original dataset. In this study, we focus on per capita income, which is sourced from total household income, and expenditure, which is derived from total household expenditure. The Oxford Scale (also known as the OECD Equivalence Scale) is used to calculate the average income and expenditure for each household member. Instead of simply dividing income and expenditure by household size (the divisor), the Oxford Scale adjusts the divisor according to the following rule:

- The first adult receives 1 point.
- Each subsequent adult is assigned 0.7 points.
- 0.5 points are given to each child.

Using the Oxford Scale as a divisor to divide household income and expenditure, the per capita income and expenditure are then calculated. Later, the experiment is continued by taking the natural logarithm form of per capita income and expenditure. The use of natural logarithms ensures that the income and expenditure distributions are more symmetrical, and it eases the building of regression models. Tables 3 and 4 display all the descriptive statistics for the variables used in this study.

Table 3. Descriptive statistics

Variable	Mean	Std Dev.
Total_Exp_01_12	3749.03	1844.23
Per_Capita_Exp	1498.32	860.39
Log_Exp	7.17	0.54
Total_Inc	6428.27	4083.59
Per_Capita_Inc	2558.87	1829.71
Log_Inc	7.64	0.63
Age	47.01	13.76
Saiz_HH	3.54	1.34
Highest_Certificate	3.12	1.60

Table 4. Descriptive statistics for categorical variable

Variable	Number of Observations, N
Sex	
Male	11895
Female	2630
Strata	
Urban	10872
Rural	3653
Region	
Centre	2222
East	2351
North	3400
South	2197
East Malaysia	4355
Ethnicity	
Bumiputera	9741
Chinese	3290
Indian	884
Others	610

3.3. Cluster Analysis

This experiment set k=5, meaning that there will be 5 clusters used to perform clustering. The optimal k is found by looking at the elbow, as shown in the scatter plot in Figure 2.

To strengthen the assumption that k=5 is the best value, silhouette analysis is performed. The experiments are repeated 7 times, for k value in the range of 2 to 8. For each iteration, average silhouette score is computed. The best silhouette score obtained is at k=5, where the score is 0.4951. Figure 3 shows the silhouette plot when k=5. It can be noticed that from all the clusters, their silhouette scores exceed the average silhouette score (denoted by the vertical dotted line). Besides that, all observations in each cluster have no negative value.

Table 5 presents the clustering results, which highlight socioeconomic factors across clusters. Cluster 3 represents the rural cluster because it includes all rural homes, whereas Clusters 1, 2, and 4 represent the urban cluster, with fewer than 10% of each cluster comprising rural households. Cluster 5 is a mixed cluster, with the majority (68.85%) hailing from cities. Clusters 3 and 5 are more likely to be multidimensionally poor because they have the lowest means for total income, total expenditure, per capita income, and per capita expenditure than any of the other clusters. Secondly, these two clusters exhibit the highest percentage of household heads without an educational certificate (32.54% from cluster 3 and 49.67% from cluster 5) and the lowest rates (7.25% from cluster 3 and 5.57% from cluster 5) of household heads with Degree/Advance Diploma certificate. It is noticeable that Cluster 3 has the most households with five or more members, accounting for 40.74% of all households, while Cluster 5 has 31.8% in this category. Clusters 3 and 5 are categorised as two distinct groups according to the k-means approach, although both clusters are multidimensionally poor. The difference is that Cluster 3 represents rural poor households, while Cluster 5 represents mostly urban poor households. Cluster 5 has closely similar household income and expenditure to Cluster 3, implying that the urban poor is a more serious problem as urban households should have higher incomes and spend more to achieve a similar quality of life. Thus, they typically lack adequate housing, facilities, and basic services due to the higher prices of these necessities in an urban setting.

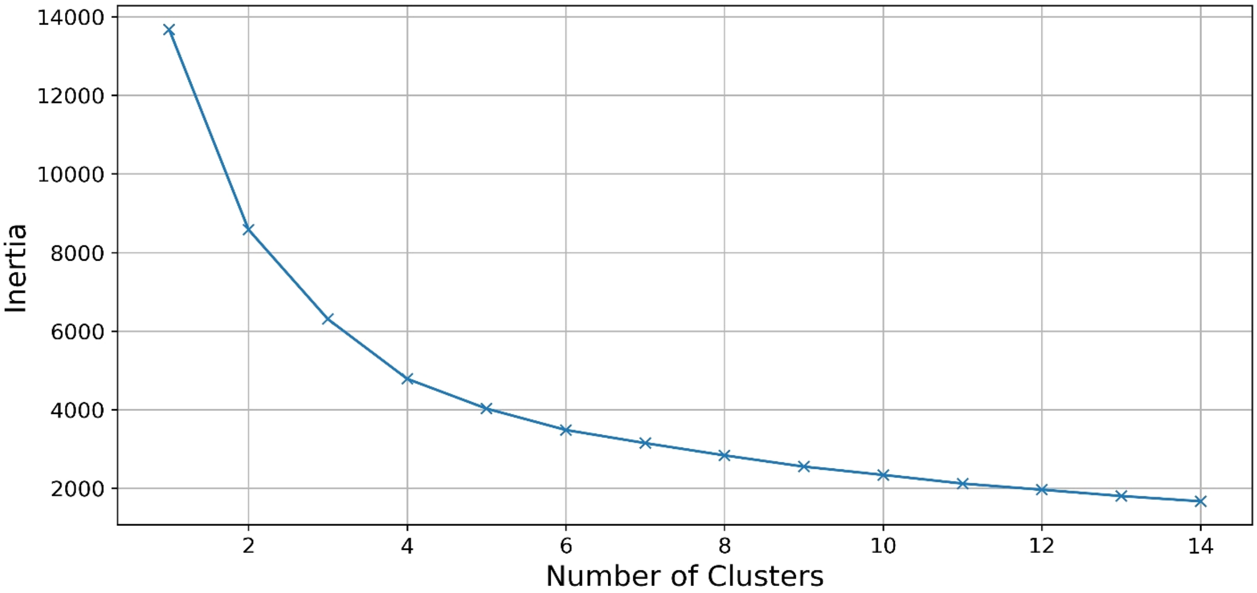

Figure 2. The Elbow Method

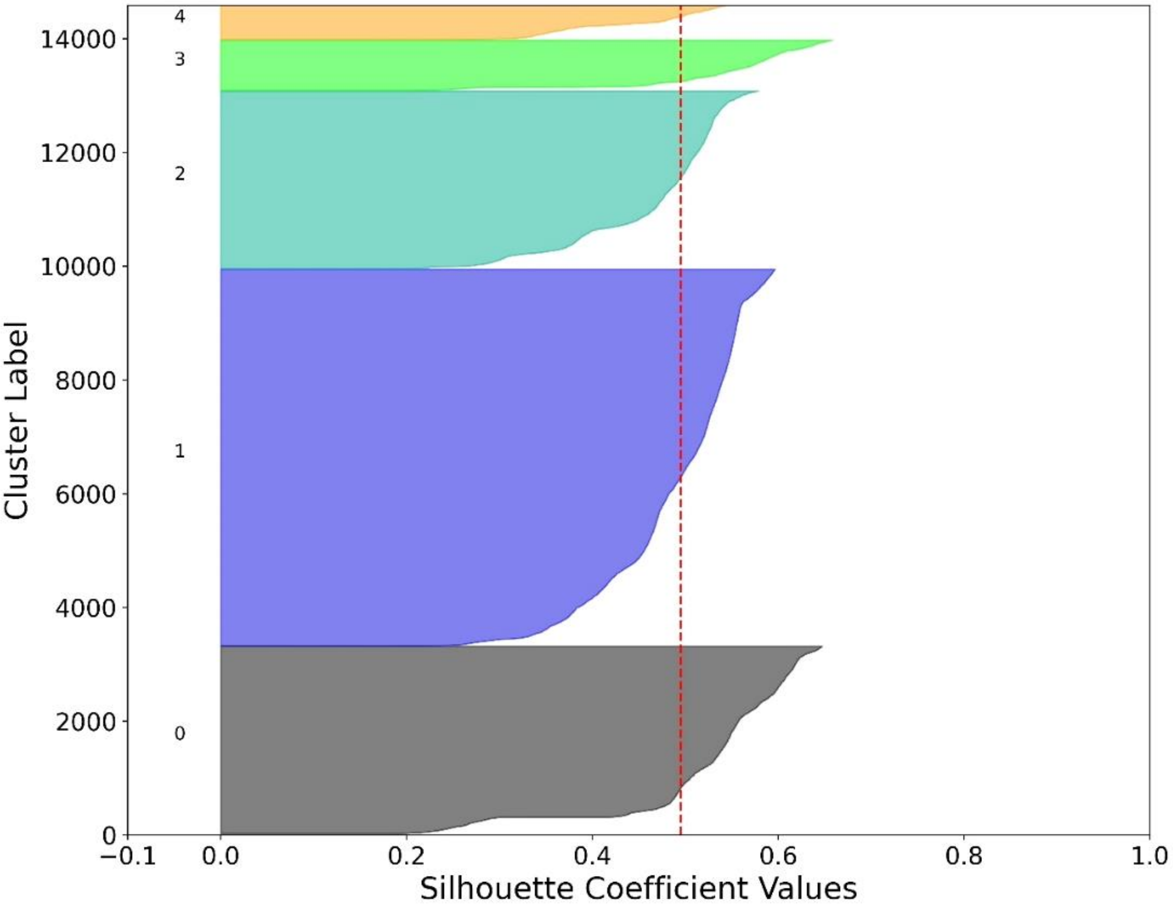

Figure 3. The Silhouette Plot

Table 5. K-means clustering result

Description	Cluster				
	1	2	3	4	5
mean of Total_Inc	6778.89	7396.75	4913.98	6709.67	4742.70
median of Total_Inc	5840.08	6408.83	3896.58	5579.10	3683.75
mean of Per_Capita_Inc	2551.42	3314.23	1862.86	2693.04	1933.38
median of Per_Capita_Inc	2123.67	2769.96	1476.63	2164.70	1583.11
mean of Total_Exp_01_12	3891.41	4371.72	2935.15	3945.72	2726.01
median of Total_Exp_01_12	3550.02	4072.31	2605.55	3575.53	2291.90
mean of Per_Capita_Exp	1467.02	1970.59	1116.33	1576.90	1131.86
median of Per_Capita_Exp	1297.22	1768.30	961.71	1360.36	930.56
mean of Age	45	51	49	47	41
median of Age	43	50	49	46	40
Highest_Certificate					
No Certificate	11.26%	21.34%	32.54%	19.80%	49.67%
PMR/SRP	9.44%	14.26%	14.29%	17.53%	10.82%
SPM/ SPMV	45.11%	37.02%	35.33%	37.10%	25.74%
STPM	3.29%	2.25%	2.98%	1.81%	2.79%
Diploma / certificate	16.88%	11.85%	7.59%	12.67%	5.41%
Degree/Advance Diploma	14.02%	13.28%	7.27%	11.09%	5.57%
Saiz_HH					
1	6.19%	11.61%	9.10%	7.35%	10.49%
2	14.29%	26.78%	16.08%	19.12%	20.00%
3	18.87%	22.71%	17.49%	20.59%	19.18%
4	21.53%	19.60%	16.59%	22.29%	18.52%
5	39.11%	19.30%	40.74%	30.66%	31.80%
Strata					
Rural	0%	8.60%	100%	6.56%	31.15%
Urban	100%	91.40%	0%	93.44%	68.85%
Number of Observations	6619	3290	3122	884	610

3.4. OLS Regression Analysis – Econometric Model

Table 6 presents the regression result using natural logarithm form of per capita expenditure (the dependent variable) and per capita income (the determinant/independent variables). The left side of the table shows the determinants, the intercept (constant), and the R^2,while the right side shows the unstandardized coefficients. Asterisks denote the statistical significance of variables. The *t*-test utilises the *p*-value to test whether an independent variable have a significant relationship with the dependent variable. As an example, using a significance level of 0.05, hypothesis testing can be carried out to determine whether household size is a significant determinant in explaining log per capita expenditure:

- H0: Household size has no effect over log per capita expenditure.
- H1: Household size has significant effect over log per capita expenditure.

Based on the result from Table 6, the *p*-value of household size (Saiz_HH) is ~0, which means the null hypothesis is rejected. In other words, at the 5 percent significance level, household size has a significant effect on log per capita expenditure. Similarly, the hypotheses for all the determinants' relationships to the dependent variable are tested through the regression model. It is found that all determinants have a significant relationship with log per capita expenditure at the 1% level of significance except the variables Age and Sex (Table 6 presents the regression result after removing these two variables). Income, educational level, strata, ethnic groups, and regions are found to have positive relationships with the dependent variable. Household size is the only variable found to have a negative impact on the log per capita expenditure.

Table 6. Regression result for linear regression (log per capita expenditure)

Variable	Coefficient
Log_Inc (Log Per Capita Income)	0.6614 ***
Highest_Certificate (Educational level)	0.0186 ***
Saiz_HH (Household Size)	-0.0598 ***
Strata (rural reference)	0.0292 ***
Bumiputera	0.3786 ***
Chinese	0.4766 ***
Indian	0.3871 ***
Others Ethnicities	0.3085 ***
Centre Peninsular	0.3642 ***
East Peninsular	0.3267 ***
Eastern Malaysia	0.2297 ***
North Peninsular	0.2545 ***
South Peninsular	0.3757 ***
Constant	1.5403 ***
R-squared, R^2	0.855 ***

Notes: *** indicate p-value <0.01, ** indicate p-value <0.05, * indicate p-value <0.1

3.5. Shapley Decomposition

The partial R^2 of the significant variables in the model of Table 7 are calculated by computing the Shapley value. Here, the Owen value is also the Shapley value or partial R^2. Panel (b) shows the general contribution from each Owen group and panel (a) shows the details of each determinant. Looking at panel (b), total income alone contributes 54.39% of R^2. Household size is the second highest, with 9.55% of R^2, follow by educational level, with 8.72% of R^2. The rest of the determinants have very minimum impact on the prediction. Noted that Shapley values will only find out the contribution of determinants, it does not point out the relationships between dependent variable and determinants.

Table 7. Shapley and Owen Value of Determinants

Variable	Owen Group	Owen values/Partial R^2
(a)		
Log Inc (Log Per Capita Income)	B1	0.5439
Highest Certificate (Educational level)	B2	0.0872
Saiz HH (Household Size)	B3	0.0955
Strata (rural reference)	B4	0.0229
Bumiputera	B5	0.0135
Chinese	B5	0.0224
Indian	B5	0.0014
Other Ethnicities	B5	0.0063
Centre Peninsular	B6	0.0214
East Peninsular	B6	0.0045
Eastern Malaysia	B6	0.0172
North Peninsular	B6	0.0035
South Peninsular	B6	0.0084
(b)		
Log Inc (Log Per Capita Income)	B1	0.5439
Highest Certificate (Educational level)	B2	0.0872
Saiz HH (Household Size)	B3	0.0955
Strata (rural reference)	B4	0.0229
Ethnics	B5	0.0436
Region	B6	0.0550
	R^2	0.848

3.6. Machine Learning - Random Forest

Random Forest comes with various hypermeters that can be set before the experiment. A good combination of hyperparameters often performs well in predicting. However, it is inefficient to attempt every possible combination manually. To deal with it, Random search function (RandomizedSearchCV from the Scikit-Learn API) is used to find the best hyperparameters. Inside the function, 5-fold cross-validation is performed too to ensure a less biased model is produced at the end. The final hyperparameters used are:

- Number of estimators/trees: 750
- Maximum depth=10
- Minimum number of samples in a leaf =4
- Minimum number of samples required to split =20

3.7. SHAP Analysis

Once the model has been trained, the testing set is then used to evaluate the model's performance and used in SHAP analysis to calculate the Shapley value. These SHAP values are used to create plots such as summary plot, dependence plot, and force plot. Figure 4 shows the SHAP summary plot. The x-axis typically represents the predictor's SHAP values. The corresponding SHAP values for that specific feature determinant are represented on the y-axis. Note that the SHAP value here refers to the dependent variable; they are having the same scale. The importance of feature determinants can be seen by looking at the y-axis of the SHAP summary plot. Log per capita income, household size, educational level, and age are variables of interest, which are discussed in Section 3.9.

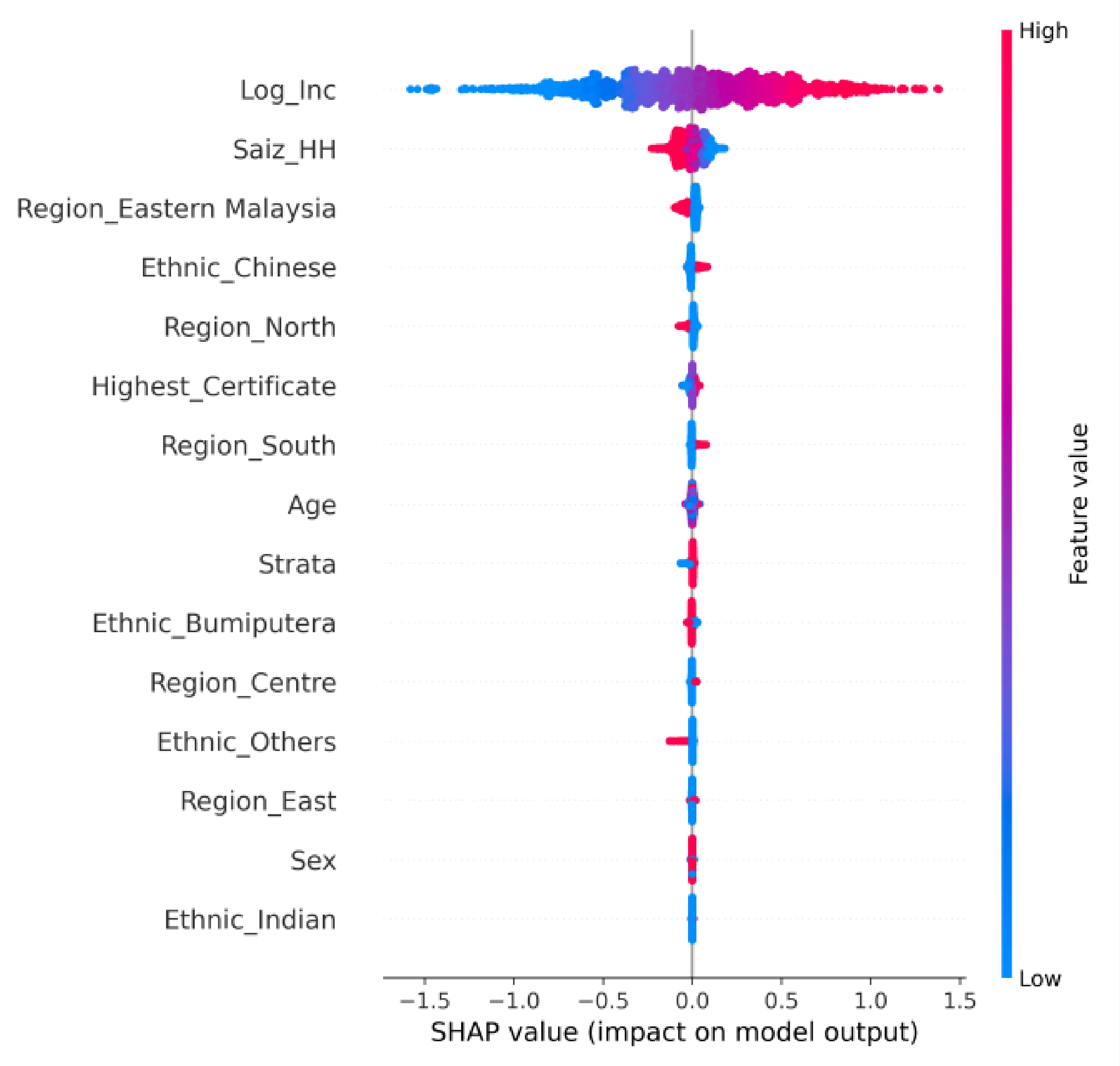

Figure 4. SHAP Summary Plot for Random Forest

The direction of the colour shift can be used to determine the relationship. For example, the colour blue on the left side of the log per capita income changes to red as it moves to the right, indicating a positive association. A positive relationship is found between educational level and log per capita expenditure. This is applicable to log per capita income too. Log per capita expenditure has a negative relationship with household size, as the colour changes from red to blue from left to right side. No clear relationship exists between log per capita expenditure and age, as denoted by the mixture of red and blue colours.

Figure 5, on the other hand, displays some significant 3-dimensional SHAP dependence plots for a few chosen determinants. By examining the trends, this plot can be used to analyse the relationship between determinants, including nonlinear relationships. Age, household size, and log per capita expenditure appear to have a nonlinear relationship in Figures 5a) and d), while Figures 5b) and c) show linear associations. The interacting or third feature value is often represented by colour in SHAP dependency plots.

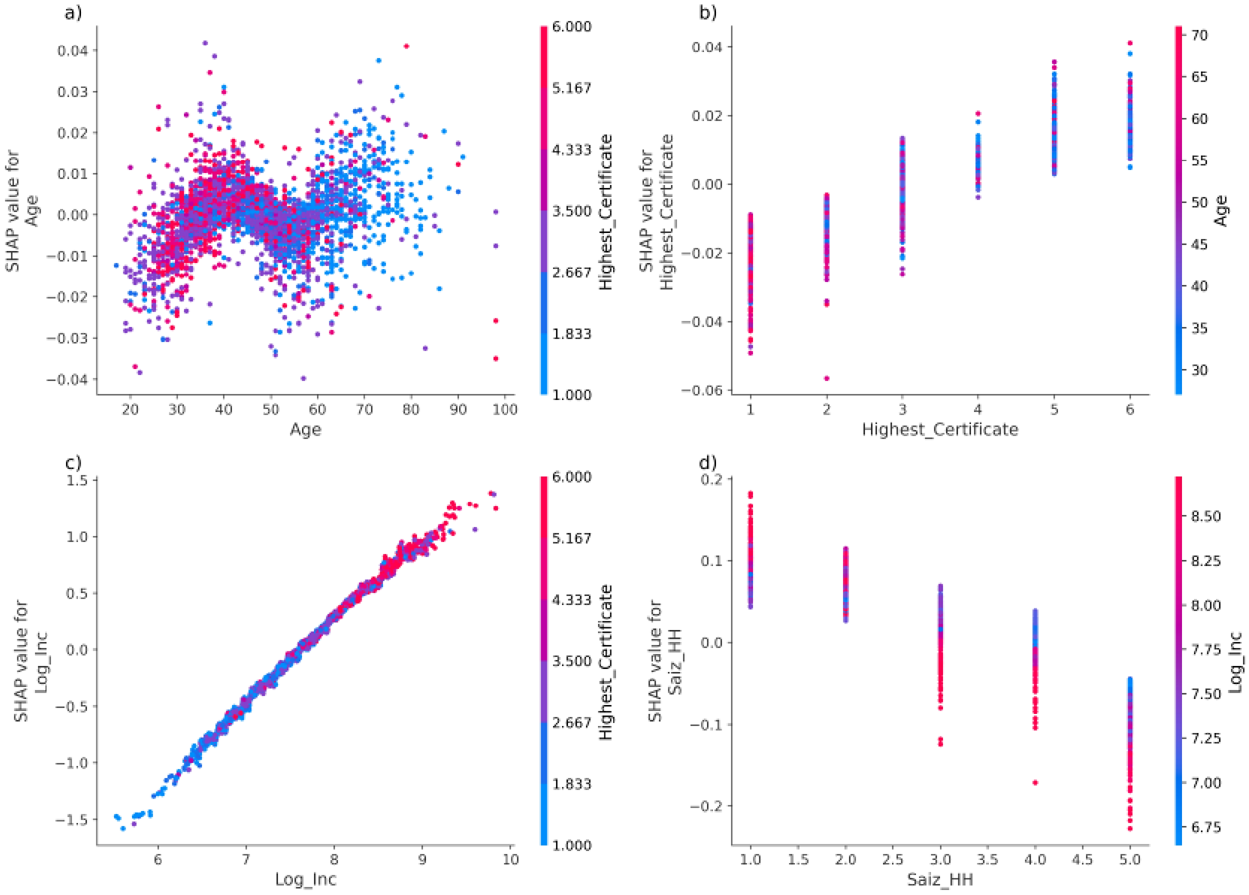

Figure 5. SHAP Dependence Plots

3.8. Models Evaluation

The R^2 and MSE values for two prediction models are presented in Table 8 for OLS (Model 1) and Random Forest (Model 2), with the testing set serving as the evaluation. Model 2: Random Forest are found to have the highest R^2 and lowest MSE scores, slightly better than the OLS model. Both models have very high R^2 and MSE scores, which means that they explain the variation very well, and do not have an overfitting problem.

Table 8. Models Evaluation

Model	Model 1: OLS	Model 2: Random Forest
R^2	0.847	0.848
MSE	0.0454	0.0453

3.9. Discussion of Findings

We begin to discuss the variables' importance to the expenditure *(Log_Exp)* first. Table 9 shows the comparison of variables' ranks for OLS and Random Forest regression. Overall, the determinants' top two rankings are similar. Income

(*Log_Inc*) and household size (*Size_HH*) are ranked as the two most powerful predictors for both models. In both models, income is ranked as the most important determinant because expenditure is generally correlated with income. Household size is ranked second in both models, as we measure household expenditure in per capita form, so an increase in a single household unit can significantly reduce per capita expenditure. OLS treats educational level (*Highest_Certificate*) as 3rd important variable, but it is only the 6th important variable in Random Forest, which unexpectedly has a lower rank than the regional variables but is still an influential factor. To this extent, both models' results are consistent with Ayyash & Sek's (2020) [16] finding, which stated that educational level and household size are the most important contributing factors toward household expenditure. Meanwhile, both models also agree that whether the household is a Chinese family can greatly determine one's expenditure power. As implied by the positive relationships in Table 6 (OLS) and Figure 5 (Random Forest), Chinese household heads generally have higher expenditures due to their higher income. This aligned with the statistics reported by DOSM Malaysia [18].

In terms of differences, in the OLS model, the top 6 variables selected are *Log_Inc* (log per capita income), *Saiz_HH* (Household size), *Highest_Certificate* (educational level), *Strata, Chinese* (Ethnics), *Centre Peninsular* (Region). All these variables come from different dimensions, e.g., income is an indicator of household economic status, and Chinese is an indicator of ethnicity. Moreover, all determinants except household size have positive relationships with log per capita expenditure. On the other hand, the Random Forest model ranks the *East Malaysia* region as the third most-important variable, while the *North Peninsular* (the Northern Region) is in 5th place. This may be because most of the states in these regions had lower mean monthly household consumption expenditures compared to the other regions, as reported in 2019 [18]. Thus, the model ranked them as significant variables in reducing the log per capital expenditure. The 6th variable is *Highest_Certificate*, which unexpectedly has a lower rank than the regional variable but is still an influential factor. From here, we can conclude that the Random Forest model focused more on regional variables. Another interesting finding is that the 2nd, 3rd, 5th determinants have negative effects on log per capita expenditure. This also suggests that the trained Random Forest model focused more on determinants that have a decreasing effect toward the expected log per capita expenditure value, as calculated by the SHAP model.

Next, we investigate the relationships between age, educational level, income, household size, and expenditure (Figure 5). Firstly, OLS found that age is statistically insignificant. Unlike the Random Forest model in Figure 5a, it shows that as age increases, the log per capita expenditure also increases. This is only true until age around 40, when, at this point, the log per capita expenditure starts to decrease until age around 60. This is believed to be the reason that the age determinant did not pass the significant test in the proposed OLS model, as age appeared in an inverted U-shaped when considering its influence toward expenditure. Chowdhury et al. (2023) [12] present similar findings too, with younger individuals tending to have a higher income and spend more, while older individuals are inclined to accumulate their wealth. However, this result contrasts with the previous findings, where expenditure has a purely positive linear relationship with age [15, 16]. Our Random Forest suggests this is the complex and nonlinear relationships that are unable to be captured by such linear model from the OLS model in this study and other studies [15, 16]. Relating to real-life situations, such an upward and downward trend is reasonable too, as young workers tend to be paid a higher salary and are more willing to spend more than older workers. Thus, this study suggests that age forms a nonlinear relationship with expenditure. As we further explore the Figure 5a), by treating the educational level as interaction feature, from age 18 to 60, most of the observations here have a higher average educational level as compared to those after age 60 (denoted by colour in the Figure 5a. Age 60 is the retirement age in Malaysia, so we could say that observations below age 60 mostly work in the formal sector, whereas observations above age 60 may be those self-employed workers in the informal sector of the economy who generally have lower education levels, as denoted by the blue colour.

Secondly, the relationship between expenditure and educational level is depicted by the coefficient (0.0186) in Table 3 (OLS) and the trend in Figure 5b (Random Forest). It is possible to characterise this positive relationship as nearly linear, but not entirely so. Comparing the educational levels of 5 and 6, their implications for expenditure are similar. In other words, the spending habits of household heads who possess diplomas, degrees, or advanced diplomas are generally comparable. This close linear relationship corresponds to the previously built clustering model (refer to Table 5). Clusters 1, 2, and 4 exhibit higher average expenditure and educational level, whereas clusters 3 and 5 have lower average expenditure and educational level. Furthermore, Ayyash & Sek (2020) [16] also noted that households with better education ought to possess greater expenditure power. In Figure 5b, too, with age serving as an interaction feature, most of the observations at educational levels 1 and 2 have a higher average age (mostly red). As educational level increases, the average age observed is lower (mostly blue). We discussed the reason for this in the previous paragraph.

Thirdly, regarding how income affects expenditure, both OLS (see Table 6, coefficient of 0.6614) and Random Forest (see Figure 5b) models agree that a positive linear relationship best represents the relationship between log per capita expenditure and log per capita income. This is not surprising since, theoretically, income and expenditure are highly correlated. As observed in Figure 5c, with educational level serving as an interaction feature, these variables are found to have perfect positive correlations; the greater any one of their values, the higher the values of the other two variables. Some observations that have a higher educational level (purple dots) but low log per capita expenditure and income (as shown in Figure 5a) can be explained by individuals who are young workers or are nearing retirement.

Lastly, investigating the coefficient of household size (-0.0598) in Table 6 and the trend in Figure 5d shows a negative relationship. Again, consistent with Ayyash & Sek (2020) [16], the larger the household size, the lower the per capita expenditure of the household head. This is not a perfect linear trend. Further analysis by integrating income as an interacting variable into household size yields more interesting patterns. If the household size is a single individual household, log per capita expenditure increases in tandem with log per capita income. In households with two members, there is no clear relationship. When household size is greater than three members, the higher the log per capita income, the lower the log per capita expenditure. This is clearly another complex relationship. The possible reason for this phenomenon is that these households prefer saving and investing rather than spending on discretionary items. Also, they may have financial goals such as saving for retirement, investing in their children's education, or building a nest egg for unexpected expenses required by household members.

On the other hand, it is crucial to point out that the expenditure pattern of those households with lower income and a household size larger than or equal to three members makes them vulnerable to being multidimensionally poor. Measuring in the absolute sense, we can take an example to explain this situation by comparing two household heads, A (with RM1000 per capita income) and B (with RM1500 per capita income). Household head A normally spends most of his/her income (say, RM800), and thus their saving per month (RM200) is far less than household head B, who spends part of his/her income (said RM600), and thus their saving per month is RM900. This overspending behaviour may come from conspicuous consumption due to low levels of human capital (in our case, educational level) typically found amongst those living in poverty [19]. This requires further analysis of the types of goods and services purchased, which is not covered in this study. One may wonder the accuracy of the example given because both the per capita income and expenditure in Figure 5d are in log form, but the log-based income and expenditure can be easily converted back by using the exponential function. Thus, our finding suggests that households that comprise three or more members with lower incomes but typically have higher expenditures require attention from policymakers.

At a glance, this study utilises both linear (OLS) and nonlinear (Random Forest) to demonstrate the relationships between socioeconomic factors and household consumption expenditure. Relying on a nonlinear random forest model allows the model to process complex relationships that exist in some determinants, achieving a more accurate and promising result. Moreover, the SHAP analysis lends a helping hand for us to visualize the trend inside a graph up to 3 dimensions, conveying our findings in a simple yet convincing manner to policymakers. To the best of our knowledge, this is the first study attempt to use the Machine Learning model followed by a SHAP model in analysing the socioeconomic variables that contribute to household consumption expenditure.

The findings of this study significantly contribute to the understanding of microeconomic dynamics within Malaysia's socio-economic landscape, offering insights into the current policy-making decision. Firstly, the observed positive linear relationship between income, expenditure, and educational level emphasises the importance of investing in education to boost economic growth and promote individual and household well-being. Secondly, the nonlinear relationship between age and expenditure highlights the need for targeted policies to address the spending habits of different age groups, particularly those above the age of 40 who may require specific support in managing their finances. Last but most importantly, the complex relationship between household size, income, and expenditure required caution and strategic policy-making.

It is essential to assist those households that comprise three or more members with lower incomes but typically have higher expenditures, particularly those with children or infants. Possible interventions can be done, including some non-cash interventions such as housing subsidies, utility bill assistance, and free health insurance. We strongly suggest the continuation of current running policies such as the RM40 Electric Rebate Programme (utility subsidy), mySalam (free health insurance), and Back-to-School Aid (cash for children in primary and secondary school), primarily targeting the group described above. It is advisable to exercise caution when it comes to maintaining direct cash subsidy programmes like i-SINAR and Bantuan Sara Hidup (BSH), given the intricate expenditure patterns observed in the aforementioned group. While direct cash transfers can provide immediate financial relief, there is a risk that households may channel them for immediate consumption instead of wealth accumulation through long-term saving and investment schemes, as Moav & Neeman (2012) [19] have stressed. Ensuring equal access to educational resources to improve human capital is one of the approaches to increasing their saving rate. Last but not least, our OLS ranking and clustering results, presented in Tables 5 and 9, also suggest the aid should be focused on urban households that exhibit identical characteristic patterns to those found in Cluster 3.

Table 9. Variables' Ranks for OLS and Random Forest Regressions

Rank	OLS	Random Forest
1.	Log_Inc (Log per capita income)	Log_Inc (Log per capita income)
2.	Saiz_HH (Household size)	Saiz_HH (Household size)
3.	Highest_Certificate (educational level)	Region (East Malaysia)
4.	Strata (Urban ref.)	Chinese (Ethnics)
5.	Chinese (Ethnics)	Region (North)
6.	Region (Central Peninsular)	Highest_Certificate (educational level)
7.	Bumiputera (Ethnics)	Region (South)
8.	Region (South Peninsular)	Age
9.	Other (Ethnic)	Strata (Urban ref.)
10.	Region (North Peninsular)	Bumiputera (Ethnics)
11.	Region (East Malaysia)	Region (Centre)
12.	Region (East Peninsular)	Other (Ethnic)
13.	Ethnic (Indian)	Region (East)
14.		Sex
15.		Ethnic (Indian)

4. Conclusion

This research focuses on household expenditure patterns in Malaysia, using 14,525 households from two dataset records: households and members. To estimate the relationship between the log of per capita expenditure and its various determinants, this study uses both a linear approach (via Ordinary Least Squares, or OLS) and a nonlinear approach (via Random Forest). In the socioeconomic field, recent studies have proved the robustness of machine learning in terms of its accuracy in predicting and complexity in handling linear and nonlinear data [10–13]. On the other hand, the traditional linear econometric model remained a popular choice in most studies due to its simplicity and standard interpretability (from coefficients). Considering this situation, we select both OLS and Random Forest models as our solution to examine the contribution of socioeconomic factors toward household expenditure by comparing how they treat the relationships between these variables. To address the black box problem inside Random Forest, we propose the SHAP model to visualise the correlations, providing valuable insights and interesting findings.

Overall, both models are powerful in predicting the consumption expenditure power of a household head, as they explain about 85% proportion of the variance (R^2) and obtain an MSE score of 0.0045 using both the training and testing sets. This also suggests both models generalise well toward unseen household data. Firstly, regarding the determinants' importance, both models suggest that income is the most important variable in explaining household expenditure. The second one is the household size. The educational level is ranked differently in the two models, with the OLS model ranking third and the Random Forest model ranking sixth. Comparing the OLS model from this study and Ayyash & Sek's (2020) [16] study, household size and educational level are two influential factors in explaining per capita expenditure. In this study, the OLS model tends to favour a diversity of determinants from different dimensions: income (log per capita income), household characteristic (Household Size), educational level (Highest Certificate), geographical location (Strata, Region) and ethnicity (Chinese). On the other hand, the Random Forest model ranked regional determinants higher in predicting per capita expenditure, which are East Malaysia as well as the North and South Peninsular regions. Although both models presented different ranking results, the difference is minor.

Furthermore, with respect to the relationship estimates gathered from both models, this work highlights that there is a positive linear relationship between household head income and educational level and their propensity to spend more, exhibiting positive linear relationships. These relationships align with the findings of Ayyash & Sek's (2020) study [16]. However, as shown in Figures 5a and 5d, nonlinear relationships indeed exist. Firstly, the larger the household, the lower the household's per capita expenditure. The contribution of each household size toward per capita expenditure varies within a boundary. As soon as we put in the per capita income, the relationship becomes obvious. The most interesting finding from Figure 5d is that if a household has three or more members, then the household heads with lower per capita incomes will spend more on basic needs and wants than those who have higher per capita incomes. Secondly, the per capita expenditure increases initially with the household head's age and then starts to decline when he/she reaches 40. The consumption expenditure subsequently decreases until age 60, which is the retirement age of Malaysia.

The research findings indicate significant implications for policy-making and interventions aimed at reducing disparities in household spending. As presented in Figure 5a, the nonlinear relationship between age and per capita expenditure highlights the need for specific targeted policies to address the expenditure patterns of different age groups, especially for those household heads before and after age 40. Looking at Figure 5d, the current policy should emphasise that those households with a size larger than or equal to 3 earn less but spend more. Subsidy assistance such as housing subsidies, utility bill assistance, and free health insurance are recommended. Direct cash subsidy programmes running right now, such as the Bantuan Sara Hidup (BSH) and i-SINAR, must establish rigorous approval processes to ensure the funds are allocated to household heads that are able to maximise their wealth through long-term investments. The priority of the assistance programmes should be to benefit urban poor households, followed by rural poor households. We contend for policies that seek to increase education levels among household heads and members to achieve higher living standards and enhanced well-being for all. In short, strategies that aim to reduce disparity in consumption expenditure should consider multiple dimensions of household characteristics.

This study reveals the possible characteristics of households that are vulnerable to being multidimensionally poor in terms of their expenditure pattern, income, age, and household size. However, it also has some limitations. Firstly, this study did not further investigate the different expenditure types. Secondly, the log per capita income itself explains more than half of the variability of both OLS and Random Forest models, so the influence of the determinants is minimized. However, as we focused primarily on decomposing the household consumption and expenditure pattern, including income is necessary. Thirdly, a wider variety of machine learning models should be considered to ensure the robustness and reliability of the findings. Finally, the findings of the OLS model and the SHAP analysis used to interpret the Random Forest model should be viewed as correlational, not as casual inferences.

5. Declarations

5.1. Author Contributions

Conceptualization, T.S.O. and Y.L.; methodology, E.L. and T.S.O.; software, E.L.; validation, T.S.O. and Y.L.; formal analysis, E.L. and T.S.O.; investigation, E.L. and T.S.O.; resources, T.S.O. and Y.L.; data curation, E.L. and Y.L.; writing—original draft preparation, E.L.; writing—review and editing, T.S.O. and Y.L.; visualization, E.L.; supervision, T.S.O. and Y.L.; project administration, T.S.O.; funding acquisition, T.S.O. All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available in the article.

5.3. Funding

The work was supported by Telekom Malaysia Research & Development under Grant RDTC/241111 (MMUE/240064).

5.4. Acknowledgements

The author would like also to thank the Department of Statistics Malaysia (DOSM) for providing the Household Expenditure Survey dataset that was essential in conducting this study.

5.5. Institutional Review Board Statement

Not applicable.

5.6. Informed Consent Statement

Not applicable.

5.7. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6. References

[1] Akyelken, N. (2020). Urban conceptions of economic inequalities. Regional Studies, 54(6), 863–872. doi:10.1080/00343404.2020.1732902.

[2] Saari, M. Y., Dietzenbacher, E., & Los, B. (2014). Production interdependencies and poverty reduction across ethnic groups in Malaysia. Economic Modelling, 42, 146–158. doi:10.1016/j.econmod.2014.06.008.

[3] Todaro, M. P., & Smith, S. C. (2020). Economic Development. Pearson Hall, London, United Kingdom.

[4] Wang, X. (2022). On the Relationship Between Income Poverty and Multidimensional Poverty in China. International Research on Poverty Reduction, 85–106. doi:10.1007/978-981-19-1189-7_5.

[5] UNDP. (2023). Unstacking Global Poverty: Data for high impact action. In Global Multi-dimensional Poverty Index 2023. United Nations Development Programme, New York, United States. Available online: https://hdr.undp.org/content/2023-global-multidimensional-poverty-index-mpi#/indicies/MPI (accessed on March 2024).

[6] Abdul Rahman, M., Sani, N. S., Hamdan, R., Ali Othman, Z., & Abu Bakar, A. (2021). A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group. PloS one, 16(8), e0255312. doi:10.1371/journal.pone.0255312.

[7] Rani, M. S. A., Nordin, S. H. B., Chin Ching Lau, Lim, S. heng L., & Siow, Z. S. (2016). Rich Debt, Poor Debt: Assessing Household Indebtedness and Debt Repayment Capacity. BNM-BIS Conference on Financial Systems and the Real Economy, 91, 153–166.

[8] Bhanoji Rao, V. V. (1981). Measurement of deprivation and poverty based on the proportion spent on food: An exploratory exercise. World Development, 9(4), 337–353. doi:10.1016/0305-750X(81)90081-4.

[9] Kumar, T. K., Mallick, S., & Holla, J. (2009). Estimating consumption deprivation in India using survey data: A state-level rural - Urban analysis before and during reform period. Journal of Development Studies, 45(4), 441–470. doi:10.1080/00220380802265207.

[10] Herrera, G. P., Constantino, M., Su, J. J., & Naranpanawa, A. (2023). The use of ICTs and income distribution in Brazil: A machine learning explanation using SHAP values. Telecommunications Policy, 47(8), 102598. doi:10.1016/j.telpol.2023.102598.

[11] Hwang, Y., Lee, Y., & Fabozzi, F. J. (2023). Identifying household finance heterogeneity via deep clustering. Annals of Operations Research, 325(2), 1255–1289. doi:10.1007/s10479-022-04900-3.

[12] Chowdhury, R. A., Ceballos-Sierra, F., & Sulaiman, M. (2023). Grow the pie, or have it? Using machine learning to impact heterogeneity in the Ultra-poor graduation model. Journal of Development Effectiveness, 1–20. doi:10.1080/19439342.2023.2276928.

[13] Zeng, Q., & Chen, X. (2023). Identification of urban-rural integration types in China – an unsupervised machine learning approach. China Agricultural Economic Review, 15(2), 400–415. doi:10.1108/CAER-03-2022-0045.

[14] Ang, W. C., & Cheah, Y. K. (2023). Inequalities in Consumption Expenditure on Pharmaceuticals: Evidence from Malaysia. International Journal of Social Determinants of Health and Health Services, 53(4), 528–538. doi:10.1177/27551938231170831.

[15] Zawiah, W., Zin, W., & Nabilah, S. F. (1998). Malaysian Household Consumption Expenditure: Rural vs Urban. Department of Statistics, Malaysia, MyStats 2015 Conference Papers, Kuala Lumpur, Malaysia.

[16] Ayyash, M., & Sek, S. K. (2020). Decomposing inequality in household consumption expenditure in Malaysia. Economies, 8(4), 83. doi:10.3390/economies8040083.

[17] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, NIPS 2017, 30.

[18] DOSM. (2020). Penemuan Utama - The Key Findings. Department of Statistics Malaysia (DOSM), Kuala Lumpur, Malaysia.

[19] Moav, O., & Neeman, Z. (2012). Saving Rates and Poverty: The Role of Conspicuous Consumption and Human Capital. Economic Journal, 122(563), 933–956. doi:10.1111/j.1468-0297.2012.02516.x.