PREDICTIVE MODELLING ON HOUSEHOLD INCOME IN MALAYSIA USING MACHINE LEARNING APPROACHES

Noryanti Muhammad
Centre for Mathematical Sciences
Universiti Malaysia Pahang Al-Sultan Abdullah
Lebuhraya Persiaran Tun Khalil Yaakob
26300 Kuantan, Pahang, Malaysia
Email: noryanti@umpsa.edu.my

Mohamad Ridzuan Shak Ibrahim
Centre for Mathematical Sciences
Universiti Malaysia Pahang Al-Sultan Abdullah
Lebuhraya Persiaran Tun Khalil Yaakob
26300 Kuantan, Pahang, Malaysia
Email: wan_ridzuan9090@yahoo.com

ABSTRACT

Household income refers to the total earnings received by all members of a household, including wages, salaries, business profits, investment returns, and other sources of income. Overall, this research looks at the causes of the rising cost of living by examining the two main factors contributing to the rising cost of living, namely the low-income factor as well as the rapid increase in living standards. To obtain the elasticity of income expenditure, a multiple linear regression was specifically used in this research on the data of the three main types of household expenditure, namely food, transportation, and housing. The elasticity of household expenditure-income was also examined according to income strata (B40, M40, and T20) as well as household location (urban or rural) to see the changes in the elasticity of expenditure-income against these variables. The goal of this paper is to find the best machine learning models for classifying household earnings using Multiple Linear Regression (MLR), Decision Tree (DT), and Random Forest (RF). To ensure the quality of the training data, several data pre-processing tasks are required, including data cleaning, feature engineering, normalization, and feature selection which consider data science cycle methodology. The correlation attribute selection has been carried out on the raw dataset. The results of this study found that the income has increased faster than the inflation rate from year to year. Furthermore, the DT and RF methods are considered the best to classify household earning. As conclusion, spending patterns and lifestyles are the dominant causes leading to the problem of the cost of living.

Keywords: Decision tree, household income, random forest, regression model

INTRODUCTION

The rising cost of living is known happened mainly in Malaysia, namely the low-income factor as well as on the rapid increase in living standards that happened in this millennial century. Everywhere around the world is impacted by these changes which impacted directly or indirectly on the community livelihood and living standard at regional level, national level and finally, at the state level. These resulted the household expenditure is known especially on the three main category, namely food, transportation, and housing, which change from time to time, depends on the current market demands particularly with the expansion of developed and developing region like in Malaysia.

By size, population, and economic level, Middle-economic Countries (MICs) throughout the world constitute a heterogeneous group. Upper middle-income economies are those with a Gross National Income (GNI) per capita between USD 4,046 and USD 12,535 (The World Bank, 2021), while lower middle-income economies are those with a GNI per capita between USD 1,036 and USD 4,045 (The World Bank, 2021). Middle income countries are home to 75% of the world's population and 62% of the world's poor. At the same time, MICs represent about one third of global GDP and are major engines of global growth (The World Bank, 2021).

As an annual population growth rate of 0.2 percent, Malaysia's predicted total population in 2022 will be 32.7 million, up from 32.6 million in 2021 (DOSM, 2022). The decrease in the number of non-citizens from 2.6 million in 2021 to 2.4 million in 2022 is the cause of the population growth rate slowing down (DOSM, 2022). This is consistent with the travel restrictions put in place by various nations when the COVID-19 epidemic spreads over the world in 2020 and 2021 (DOSM, 2022). The number of citizens grew from 30.0 million in 2021 to 30.2 million in 2022, with a growth rate that dropped from 0.8% to 0.7% over that time (DOSM, 2022).

The pattern of population distribution is known by age group, gender, state, education level, occupation, and urbanization are affected by changes in Malaysia's demography and socioeconomic conditions, which are brought on by changes in the total population composition of the nation. In any nation, having favourable socioeconomic conditions is vital for a fulfilling life. Planning and development for the socioeconomic sector must be done in an organized and purposeful manner. The entire population should benefit from prosperity and happiness. The income level and distribution of household expenditure must reflect the strength and expansion of the economy in order to accomplish the aim. Therefore, it is crucial to use the right measures for assessing and monitoring issues like employment, poverty, and income inequality.

An individual's socioeconomic position or wellbeing can be measured in part by their income. Using a household survey or administrative records are the two approaches that can be used to gather information about income (Ursuna K., 2019). On the

other hand, it is practical to ask households through surveys about their income. Data on household income are available from the Household Income and Basic Amenities Survey (HIS & BA), which was carried out in Malaysia by the Department of Statistics Malaysia (DOSM).

Household income is known as the total income earned by all members of the home. The household income for a single dwelling is calculated by adding the gross income of each adult resident of the home who is 15 years of age or older, whether they are related. The key elements that have a substantial impact on household income are location (rural or urban), education, gender, employment history, and age. According to the International Monetary Fund (2001), many developing nations experience the worldwide problem of rising living expenses. Affluent nations like the United States and Japan are also dealing with the issue of their citizens' rising cost of living. For instance, there are significant differences in purchasing power between cities in the United States (Department of Economic and Social Affairs, 2020). Despite having some of the highest costs in the country for commodities, New York City has some of the middle to ninth top wages in the country. As a result, New York City's purchasing power levels in 2014 were among the lowest of all American cities. The same circumstance has also given Philadelphia a poor ranking in the United States' city-by-city purchasing power chart (Sen and Adam, 2017). The scenario indirectly explains that the increase in the price of goods is much greater than the increase in wages of workers in the area. Malaysia also addresses the issue of a high cost of living in an indirect manner. According to Worldwide Cost of Living Survey (2019), Kuala Lumpur is the 85th most expensive city in the world in terms of cost of living. Malaysia's growing cost of living is a continuous issue that has existed for a long time and the problem has gotten worse because of the huge rise in housing expenses over the previous 10 years (Business Today, 2019).

Based on RHB Research Analyst (2018), the current situation is different from the low-interest rate cycle around 2012 to 2013 which saw the House Price Index (HPI) increase between 11 to 13 percent during the period. However, the HPI growth declined to -0.7% in the third quarter of 2021 compared to 3.3% in 2018 (Bank Negara Malaysia, 2018). The research was done to identify important variables that affect household income in Malaysia in the year 2021. The study proposed linear regression model of logs with interaction variables to predict the household income in Malaysia. A study on the effect of income increase on spending patterns was done by analysing the level of household expenditure-income elasticity (James Petras, 2007). The analysis of expenditure elasticity provides an idea of the extent to which an item is an important component as well as the level of household needs based on the income owned. Therefore, this study has examined the effect of increased income on expenditure for households according to income strata B40, M40 and T20 in urban and rural areas by using the linear regression model of logs with interaction variables. Increased income is among the main factors to the improvement of living standards (James Petras, 2007), then could be able to see a direct relationship between living standards, income, and expenditure. If spending increases faster than income, this gives an indication that the standard of living of households has increased, thus causing their spending to be higher, and not simply because of rising prices of goods.

Household income is one of the important factors or sectors which are important to indicate our economic level. As technology increases, we believe that the household income might affect and need to be investigated in detail. The common family earnings of Malaysia distended with the help of Rancangan Malaysia Ke-11 (RMK-11) to RM5,900 a month, compared to RM5,000 in 2012 (Economy of Malaysia, 2021). The unit financial gain “Estimates and Incidence of Impoverishment Report, Malaysia” (2020) which incorporates the findings of COVID-19 result study on household income in Malaysia for the twelve months in year 2020. This record to support consists of the estimation on prevalence of impoverishment in Malaysia with inside the equal twelve months (DOSM, 2019).

The government is more focusing on identifying the level of household income with different levels for the effective solution. This low-income cluster remains significantly vulnerable to economic shocks additional as it may increase among the worth of living and mounting cash obligations (The World Bank, 2021). Therefore, in this research, the researcher wanted to develop a new model on household income at Malaysia which consider up-to-date variables such as technology expanse might be the trigger or may contribute to increase in household income in Malaysia by using linear regression model of logs with interaction variables.

Machine learning (ML) approaches are essential for analyzing household income in Malaysia due to their ability to handle high-dimensional, nonlinear relationships and improve predictive accuracy. Many other researchs use ML to analyse data (Al-Maula et al, 2025, Lewaaelhamd & Iskander, 2025, Kumar, 2025, Nkurunziza et al, 2025) Unlike traditional regression models, ML techniques such as decision trees, random forests, and deep learning can automatically identify significant variables, recognize complex patterns, and adapt to changing economic conditions. Additionally, ML can integrate diverse data sources, including technological expansion indicators, to assess their impact on income levels. By leveraging ML, the study can generate more precise income predictions, enhance policy decision-making, and provide targeted solutions for different income groups (B40, M40, T20) and regions (urban vs. rural). The adaptability of ML ensures continuous learning from new data, making it a powerful tool for understanding and addressing Malaysia's rising cost of living.

METHODOLOGY

In this section, the data science cycle methodology is used. To achieve the objectives of the study, evaluation on the household income features in determining the current effect variables are studied. The systematic evaluation process is done on the proposed machine learning algorithm based on six stages and shown as in Figure 1.

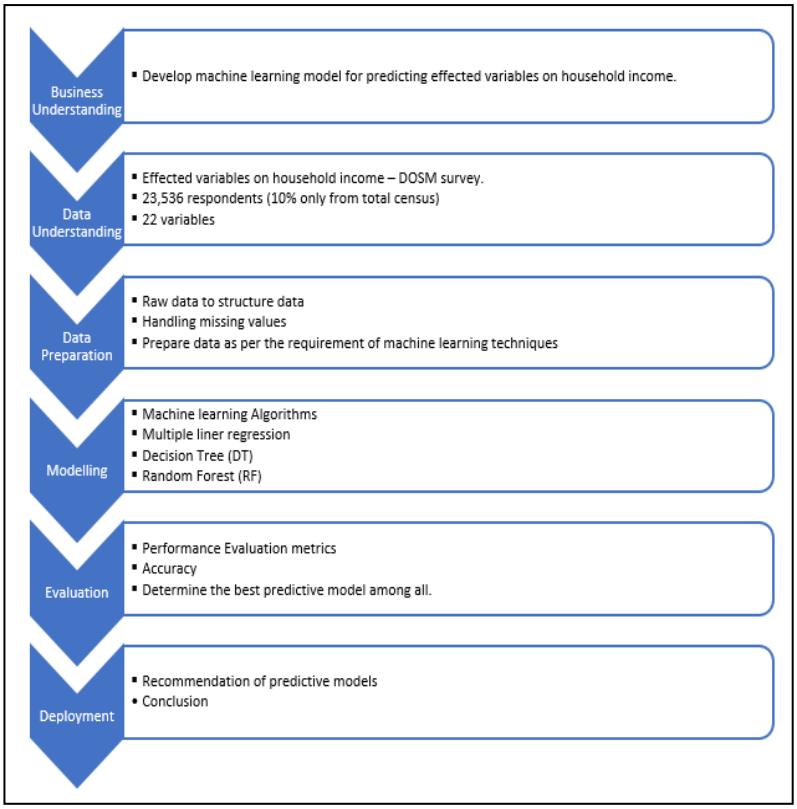

Figure 1: Data Science Cycle Methodology

Based on the Figure 1, the overall household income situation at Malaysia has been explore and understand. The data was taken from Department of Statistics Malaysia (DOSM) for the year 2016, which consists of 22 variables. Basically the population of the study is Malaysian. Then, data preparation has been done which considering the structure of the data, handling missing values, features selection, and etc, before it been used to be modelled. As mentioned in this study, machine learning approaches are used, which are Multiple Linear Regression (MLR), Decision Tree (DT) and Random Forest (RF).

In the MLR, the impact of numerous explanatory variables on a particular result is considered. The MLR model for dependent variable of with k predictor variables can be written as in equation (1).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon i \quad (1)$$

where is the dependent variable, β_0, β_1, …, β_k are coefficients of regression to be estimated with respect to observations, x_{i1}, …, x_{ik} are explanatory variables and is the error term. A statistical metric called the coefficient of determination (R^2) or Adjusted R^2 is used to assess how much of the variance in the result can be accounted for by the change in the independent variables. Even though the predictors may not be connected to the outcome variable, adjusted R^2 always rises when more predictors are included in the MLR model.

A DT is a supervised learning tree-structure based support tool for both regression models. The value of the output variable can potentially be estimated using an upside-down tree-like structure (with leaves at the bottom) by stratifying portions of predictor space (no overlap). Decision trees (DT) often have an inverted structure, and internal nodes are what separate a space into sub-spaces. The elements that connect the nodes are called branches. A decision support tool known as a decision tree employs a tree-like model to represent options and their potential outcomes, including utility, resource costs, and chance event outcomes. One technique to show an algorithm that solely uses conditional control statements is to use this method.

While RF feature techniques is known were created to lower the variance of the model while keeping a low bias. The forests produce several decision trees concurrently while selecting a random assortment of the split method's attributes. Due to the greedy feature selection process used by nodes, it is possible to decorate distinct decision trees that can be strongly coupled when fitted with bootstrapped data. By bootstrap aggregation, trees are fitted to random samples of training data. Throughout tree training, the model determines the contribution of each feature to reducing the variance. The ranking of the attributes in terms of importance is possible for a forest using the average of the variance decrease for each characteristic. The fact that picking features based on decreasing variance promotes variables with more categories should be underlined. Additionally, when there are related qualities, the model automatically minimises the importance of the others and selects any one of them as a predictor without clearly favouring it.

In order to assess model accuracy and avoid overfitting, which occurs when a model "performs well on the training set but is unable to generalise to new data," cross-validation techniques are utilised. K-fold cross-validation (for training) and holdout approaches are used to assess the performance of the three machine learning algorithms. The dataset is split into two sets using the hold-out method: 20% for testing and 80% for training. The dataset is split into 5 equal-sized divisions using a 5-fold cross-validation procedure to prevent overfitting. The same model is trained five times, using the other four pieces for training and one component for testing each time. The five findings are averaged to get a single estimate. The Root Mean Squared Error (RMSE),

Mean Absolute Error (MAE), and R2 were calculated to examine the divergence between actual and projected household important variable data for evaluating the accuracy of models and inter-model comparisons.

RESULT AND DISCUSSIONS

The model was first run using MLR model. The initial MLR model showed the problem of multicollinearity and non-significance of certain predictor variables. The checking procedure was first focused on multicollinearity and then evaluated predictor variables significance. Besides, predictors that were not significant are also discarded since they do not contribute to the model. The MSE of the final model is the main concern of this study to see the effectiveness of the model in reducing error rates and variation of household income. The MLR final model consists of 5 significant explanatory variables and has RMSE= 0.29. The significant model is shown in Table 1.

Table 1: Model Validation Machine Learning Method

No	Algorithms	RMSE	MAE	Adjusted R^2
1	Multiple Regression Model	0.29	0.082	0.99
2	Random Forest	0.001	1.1-6	0.97
3	Decision Tree	0.00	0.00	1.00

The dataset used in this study is subjected to a variety of machine learning techniques, with DT providing the greatest accuracy (95.20%) when used for classification. Random Forest (RF) model has the second-highest accuracy, with 93.20%. the use of a MLR classifier, which has a 100% accuracy rate. Three distinct datasets are used, then the accuracy of the algorithms is compared. The model's household prediction accuracy and precision are improved which shown in Table 2.

Table 2: Comparison Model Validation with Machine Learning Method

No	Algorithms	RMSE	MAE	Adjusted R^2	Accuracy
1	Multiple Regression Model	0.29	0.082	0.99	100.00%
2	Random Forest	0.001	1.1-6	0.97	93.20%
3	Decision Tree	0.00	0.00	1.00	95.20%

CONCLUSIONS

There have been numerous academic debates and policy discussions about household income, its shifting nature, and effects, how to define and measure household income for policy purposes, and the various policies and programs that have been put in place to address it. In this paper, this research does not provide a perfect model that can predict income from individual characteristics. Instead, it demonstrates the applicability of these methods used in economic research and the potential for the development of such models by those wishing to include large amounts of household income data. These implications have far-reaching implications that, if applied ethically, can have positive social benefits. In this paper, DT is one of the ways to analyse the data since it is known to be able to handle a range of data types. While RF was used to automatically pull out 80% of the training data from the dataset. Each tree was created individually, and utilizing the test dataset and the mean value, hence, the final prediction is accurate. It thus displayed best-in-class performance in terms of negligible generalization errors. As the household income dataset utilized in this study had both continuous and categorical characteristics, which categorized the income to increase throughout the year, RF seemed to be a good classifier to test the model. To improve the findings of this study, the next study examines in more detail some of the possibilities that have been mentioned, such as income differences between B40, M40 and T20 quality to be further expanded in Malaysia.

ACKNOWLEDGMENT

The authors extend their gratitude to the University of Malaysia Pahang Al-Sultan Abdullah (UMPSA) and Persidangan Kependudukan Kebangsaan 2023 (PERKKS 23) organizer (United Nations Population Fund) for the financial support, as well as for the use of UMPSA facilities. Appreciation is also extended to Lembaga Penduduk dan Pembangunan Keluarga Negara (LPPKN) for their insightful and experienced feedback throughout the research process. Additionally, the authors thank the reviewers for their valuable suggestions, which have greatly contributed to the improvement of this paper.

REFERENCES

Al Maula, S. F., Setiawan, N. A. D., & Pusporani, E. (2025). MODELING HOUSE SELLING PRICES IN JAKARTA AND SOUTH TANGERANG USING MACHINE LEARNING PREDICTION ANALYSIS. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, *19*(1), 107-118.

Babbie, E. R. (2010). The Practice of Social Research (12th ed.). Belmont, CA: Wadsworth Cengage.

Bank Negara Malaysia, (2018). Budget Speech 2018. Kementerian Kewangan Malaysia. Percetakan Nasional Malaysia Berhad

Blaxter, L., Hughes, C. & Tight, M., (2001). How To Research (2nd Ed.). Philadelphia, USA: Open University Press.

Department of Statistic Malaysia, (2020). Pendapatan Dan Perbelanjaan Isi Rumah M40 Dan B40 Mengikut Negeri. Disediakan oleh Shahriman Haron, Jabatan Perangkaan Malaysia,

DOSM/BPHPP/3.2020/Siri.https://www.dosm.gov.my/v1/uploads/files/6_Newsletter/Newsletter%202020/DOSM_BPHPP_3-2020_Siri_28.pdf

Department of Statistic Malaysia, (2022). Malaysia's Human Development Index (HDI). Department Of Statistic Malaysia Newsletter, DOSM/BPTMS/1.2022/Series 27. https://www.dosm.gov.my/v1/uploads/files/6_Newsletter/Newsletter%202022/DOSM_BPTMS_1_2022_Series%2027_compressed.pdf

International Monetary Fund (2001). Global Trade Liberalization and the Developing Countries.https://www.imf.org/external/np/exr/ib/2001/110801.htm

Kumar, M. (2025). Price Prediction Using Machine Learning Approaches. In *Proceedings of 4th International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication: MARC 2023, Volume 2* (p. 279). Springer Nature.

Lewaaelhamd, I., & Iskander, M. G. (2025). Poverty prediction using machine learning models: Insights from HICES survey in Egypt. *Statistics, Optimization & Information Computing*.

Nkurunziza, F., Kabanda, R., & McSharry, P. (2025). Enhancing poverty classification in developing countries through machine learning: a case study of household consumption prediction in Rwanda. *Cogent Economics & Finance*, *13*(1), 2444374.

Sen, E., & Adam, S., (2017). Regional Spotlight: Purchasing Power Across the U.S. Federal. Reserve Bank of Philadelphia.

Sekaran, U., (2003). Research Methods for Business: A Skill-Building Approach. 4th Edition, John Wiley & Sons, New York.

The World Bank (2021). Overview The World Bank on Malaysia. https://www.worldbank.org/en/country/malaysia/overview#1

Ursuna Kuhn, (2019). Measurement Of Income in Surveys. FORS Guide No. 02, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. doi:10.24449/FG-201900002