Asia-Pacific Journal of Information Technology and Multimedia 14(1) 2025: 40 – 60
https://doi.org/10.17576/apjitm-2025-1401-03

The Random Forest Algorithm for Modelling the Overspending Behaviour of Malaysian Households Income Class

Algoritma Random Forest untuk Pemodelan Tingkah Laku Perbelanjaan Berlebihan dalam Kelas Pendapatan Isi Rumah di Malaysia

Liyana Ihkwani Abdul Latif[1], Azuraliza Abu Bakar[2], Zulaiha Ali Othman[3], Mohd Suhaidi Abdul Rais[4], Mazniha Berahim[5]*

[1,2,3]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

[4]Department of Statistics, Federal Government Administrative Centre 62514 Precinct 1, Wilayah Persekutuan Putrajaya, Malaysia

[5]Center for Diploma Studies, Universiti Tun Hussein Onn Malaysia (Pagoh Campus), KM1, Jalan Panchor, 84600 Muar, Johor, Malaysia.

**Corresponding author: Azuraliza Abu Bakar (azuraliza@ukm.edu.my)*

Received 11 October 2024
Accepted 22 May 2025, Available online 30 June 2025

ABSTRACT

Overspending is a typical financial behaviour that can affect individuals across all income levels, but it tends to impact those with lower incomes significantly. Research has shown that low-income individuals are more likely to experience financial hardship as a result of overspending. Previous studies in socio-economic analytics have demonstrated the potential of machine learning as a predictive model. This study proposed the use of the Random Forest method to build a predictive model of overspending behaviour among Malaysian households in the B40, M40, and T20 income groups. The model was developed using the household income and expenditure data from the survey conducted by the Department of Statistics Malaysia (DOSM) in 2016. The original dataset comprises three databases containing 1.5 million records of head and household members. These databases were integrated into a single dataset with 14,551 household records and 25 parameters, including 13 demographic factors and 12 categories of household expenditure. The Random Forest algorithm achieved the highest accuracy compared to other well-known machine learning methods. Its predictive attributes were compared with the household expenditure reports from DOSM for 2016, 2019 and 2022. The overspending attributes identified from the 2016 data were consistent with

expenditure patterns in 2019 and 2022, suggesting that the proposed model can effectively predict future spending items. This study provides valuable insights into household spending and overspending behaviour and highlights the potential for further research in socio-economic analytics.

Keywords: financial behaviour, feature selection, random forest, classification, overspending

ABSTRAK

Perbelanjaan berlebihan adalah tingkah laku kewangan yang biasa yang boleh memberi kesan kepada individu dari semua peringkat pendapatan, namun ia boleh memberi impak yang besar kepada mereka yang berpendapatan rendah. Kajian telah menunjukkan bahawa individu berpendapatan rendah lebih cenderung menghadapi kesulitan kewangan akibat perbelanjaan berlebihan. Penyelidikan terdahulu dalam analitik sosioekonomi telah menunjukkan kemampuan pembelajaran mesin sebagai model ramalan. Kajian ini mencadangkan kaedah Random Forest untuk membina model ramalan tingkah laku perbelanjaan berlebihan dalam kalangan kelas pendapatan isi rumah B40, M40, dan T20 di Malaysia. Model ini dibangunkan menggunakan data pendapatan dan perbelanjaan isi rumah yang diperoleh daripada tinjauan yang dijalankan oleh DOSM pada tahun 2016. Set data asal terdiri daripada tiga pangkalan data dengan 1.5 juta rekod ketua dan ahli isi rumah. Kami menggabungkan pangkalan data tersebut menjadi 14,551 rekod isi rumah dengan 25 parameter, termasuk 13 parameter demografi dan 12 pengelasan data perbelanjaan isi rumah. Algoritma Random Forest menghasilkan ketepatan tertinggi berbanding dengan kaedah pembelajaran mesin lain yang terkenal. Atribut Random Forest dibandingkan dengan laporan perbelanjaan isi rumah oleh DOSM bagi tahun 2016, 2019, dan 2022. Atribut perbelanjaan berlebihan yang diperoleh daripada model yang dicadangkan menggunakan data 2016 adalah konsisten dengan perbelanjaan masa depan dalam data 2019 dan 2022, menunjukkan bahawa model ramalan yang dicadangkan boleh digunakan untuk meramalkan item perbelanjaan masa depan. Kajian ini menemui pandangan menarik dalam set data yang boleh digunakan untuk memodelkan tingkah laku perbelanjaan dan perbelanjaan berlebihan dalam kalangan isi rumah. Ia membuka kawasan penyelidikan baru dalam analitik sosioekonomi.

Kata kunci: tingkah laku kewangan, pemilihan ciri, Random Forest, pengelasan, perbelanjaan berlebihan

INTRODUCTION

Financial literacy is essential for achieving financial well-being, meeting daily needs, planning for the future, and making informed financial decisions. Assessing financial literacy levels across socio-economic groups can support the development of targeted financial education programs. Financial literacy is closely linked to financial behaviour, which encompasses meeting basic needs, planning, monitoring expenses, and cultivating responsible spending and saving habits. Improving financial literacy can lead to better financial behaviours, reducing the

risk of bankruptcy, avoiding high-risk financial decisions, and enhancing overall spending habits. It also positively influences financial attitudes and saving and spending practices, making it a key factor in enabling individuals to participate actively in the economy. Financial behaviour is shaped by various factors, including psychological and emotional influences, cultural and social norms, and access to financial resources and education. Individuals who are financially literate and have access to adequate resources are more likely to adopt behaviours that promote financial stability. Overspending is a type of financial behaviour that often leads to debt and financial instability. While it affects individuals across all income levels, those with low incomes are particularly vulnerable due to limited financial resources and lower levels of financial education. Overspending can stem from several causes, including emotional spending, lack of financial literacy, and present bias. It occurs when individuals spend more than they earn, leading to excessive debt and compromised financial stability. Overspending is a widespread issue with serious consequences, such as increased stress, depression, and a reduced ability to achieve long-term financial goals (Zou, Peng, & Luo, 2015). Several factors contribute to overspending. For example, individuals with low financial literacy may struggle to understand the long-term consequences of their financial decisions. Likewise, emotional spending, such as making impulsive purchases in response to stress or sadness, can lead to overspending.

Using machine learning (ML) to predict overspending behaviour can offer more detailed insights into the differences and similarities across various income groups. Combining ML with parametric models can help identify individuals with low financial literacy at risk of financial vulnerability. Previous studies have employed ML techniques such as Support Vector Machines, Decision Trees, and Artificial Neural Networks to predict financial literacy (Lusardi & Mitchell, 2014; Huang et al., 2013; Lin et al., 2012). In Malaysia, research applying ML to model spending behaviour and financial burden across income groups can be found in studies by Bakar et al. (2021) and Othman, Abu, Sani, & Sallim (2020). The predictive models developed through this research can assist policymakers and financial institutions in designing targeted financial education programs and provide necessary support and resources. Additionally, they can empower low-income individuals to take proactive measures such as budgeting and debt reduction to improve their financial behaviour.

This study aims to develop an overspending classification model using the Random Forest method. The model predicts overspending factors across different income classes in Malaysia and identifies key expenditure items contributing to overspending within these groups. While most studies on spending behaviour and financial literacy rely on survey data, this study utilises administrative data from the 2016 Malaysian Household Income and Expenditure Survey (HIES), provided by the DOSM. The key contributions of this paper are as follows:

1. The Household Income and Expenditure dataset, combined with the Classification of Individual Consumption according to Purpose (COICOP) data, is a novel approach used in ML modelling.
2. Using the Random Forest method, the proposed overspending classification model achieves a high predictive accuracy of 92.1%.

3. The overspending features identified from the 2016 HIES data are consistent with spending behaviour observed in the 2019 and 2022 datasets, indicating the predictive model is valid over time.

This paper is organised into five sections. The following section presents related work. The materials and methods section details the data and modelling process. This is followed by the results and analysis. Finally, the paper concludes with key findings and implications.

LITERATURE REVIEW

In the 10th Malaysian Plan, the government outlined the classification of income groups based on household income levels: the Top 20% (T20), Middle 40% (M40), and Bottom 40% (B40). As of 2019, the income threshold for the B40 group, which includes 2.91 million households, is RM4,849 and below. The M40 group, comprising 2.91 million households, has an income range of RM4,850 to RM10,959. Meanwhile, 1.46 million households fall into the T20 group, with incomes exceeding RM10,960. Household purchasing power is primarily determined by income, which directly correlates with expenditure. In 2019, Malaysian citizens experienced a 4.2% increase in income and spending, with disposable income rising by 4.4%. Some states saw income growth outpace expenditure growth, while others experienced the opposite trend. Malaysian households allocate 68.7% of their disposable income to consumption expenditure. On average, households have a surplus income of RM2,118 and financial commitments averaging RM3,612, typically assigned to debt repayment and investment. Notably, 30.4% of households earn less than RM4,000 per month, with 24.8% falling within the RM2,000 to RM3,999 range. It indicates that household purchasing power in Malaysia still has room for improvement, as 56.2% of households report monthly spending below RM4,000 (DOSM, 2020).

Behavioural finance theory offers valuable insights into why some individuals overspend and how such behaviour can be addressed. For instance, individuals may exhibit present bias, leading them to prioritise immediate gratification over long-term financial goals. Emotional factors such as excitement, anxiety, or guilt can also influence spending behaviour, prompting impulsive purchases or overspending in response to external pressures (Zainal et al., 2012).

Overspending is a typical financial behaviour that can affect individuals across all income levels, but it tends to significantly impact those with low incomes. Research has shown that low-income individuals are more likely to experience financial hardship from overspending, due to a combination of factors such as limited financial resources, high debt levels, and restricted access to financial services and education (Zainal et al., 2012). Overspending is particularly problematic for low-income individuals because they often lack sufficient financial buffers. When overspending, they are more likely to exhaust their savings, accumulate debt, and face ongoing financial instability.

Another factor contributing to overspending among low-income individuals is the lack of financial education and resources. Research has shown that low-income individuals are less

likely to have access to financial services such as banking and investment, and may lack the financial knowledge and skills needed to make informed decisions. It can lead to poor financial choices, including overspending and increased financial hardship. To address this issue, policymakers and financial institutions can play a crucial role by promoting financial literacy and improving access to financial education and resources for low-income communities.

Data analytics is widely recognised as a powerful approach for uncovering meaningful insights from large datasets. It supports various tasks such as classification, clustering, prediction, diagnostics, and anomaly detection. Several well-known ML methods used for prediction include Regression, Artificial Neural Networks, Support Vector Machines, and Decision Trees.

A study on Latin American ADRs (LAADRs) and Latin American banks (LABANKs) investigated the impact of corporate governance and accounting parameters on the efficiency of the two industries. Logistic regression is conducted to identify the key explanatory variables, with Tobin's Q and DEA technical efficiency indicators used as the dependent variables for LAADRs and LABANKs, respectively. The regression model included a dummy variable for industry sectors and calculated efficiency indicators for each country. The study employed the AdaBoost method to classify stocks and banks as either above or below the median based on market value and efficiency. Bagged boosting was applied to AdaBoost to evaluate the stability of the results. The odds ratios from the logistic regression further confirmed the significance of the main parameters in determining efficiency (Creamer & Freund, 2004).

Another study by Huang et al. (2008b) used survey data from 1,010 participants to model financial decisions related to credit cards, loans, and superannuation, employing Support Vector Machine (SVM) and Backpropagation Neural Network (BPNN) methods. The SVM utilised Gaussian functions as the kernel, while the BPNN applied the Conjugate Gradient method to minimise the mean squared error. The results showed that SVM outperformed BPNN in all cases, achieving an overall generalisation performance of 93%. The study suggests that SVM captures the underlying relationships between inputs and outputs better. Using the Artificial Neural Network method, Sood and Bhushan (2017) developed a Financial Literacy Prediction model (FLPNN). The dataset is based on primary survey data from 516 salaried individuals working in Himachal Pradesh. The FLPNN model demonstrated good sensitivity and specificity, achieving a total accuracy of 75%. The ROC curve had an area under the curve (AUC) of 80%, indicating strong discriminatory power.

The study by Levantesi and Zacchia (2021) used data from the Bank of Italy’s 2017 survey to investigate financial literacy (FL) and financial inclusion among Italian adults. The focus was on the knowledge component of FL, specifically the understanding of basic financial concepts. A composite FL index was used to evaluate respondents’ financial knowledge, categorising them into two groups: those with higher financial literacy and those with lower levels of financial education. Demographic factors such as gender, age, education, household composition, and employment status were considered, along with financial parameters including financial behaviour and attitudes, to identify the key determinants of higher financial

literacy among Italian adults. Personal financial behaviours such as a propensity for pension planning and high self-assessment of financial knowledge were also included to develop a model for estimating the main determinants of financial literacy in Italy.
Another study employed Random Forest regression techniques to investigate customer retention and profit outcomes. Using this method, the researchers analysed parameters such as past customer behaviour, observed customer diversity, and several other factors. The analysis was conducted on a real-world sample of 100,000 customers obtained from the data warehouse of a large financial services company in Europe. The findings indicated that the Random Forest technique provided a better fit for sample estimation and validation than linear regression and ordinary logistic regression models (Lariviere & Vandenpoel, 2005).

The study by Othman et al. (2020) employed ML techniques to identify overspending patterns and contributing factors among Malaysian household income classes B40, M40, and T20. The data, obtained from the Department of Statistics Malaysia (DOSM) in 2016, consisted of 14,451 records and included 12 parameters: number of households, area, state, strata, race, highest educational certificate, marital status, gender, housing, income, total expenditure, and income category. The study utilised various ML algorithms, including Decision Tree, Naïve Bayes, Neural Networks, Support Vector Machines (SVM), and Nearest Neighbour, to determine the parameters influencing overspending. The results showed that SVM achieved the highest accuracy at 89.17%. The six main factors influencing overspending behaviour were state, race, income, number of households, and income category. However, the study did not consider the items purchased that led to overspending.

The study by Bakar et al. (2021) employed machine learning (ML) methods to classify the financial burden among Malaysian household income classes. The researchers considered the number of household members and the relationship between the head of household and household members as primary indicators for assessing financial burden risk. The dataset from the Department of Statistics Malaysia (DOSM) covered rural and urban areas across all states. The final dataset consisted of 14,838 cases and 14 parameters, compiled from the integration of three databases: 1,058,574 household expenditure records, 64,091 cases across 14 categories of household members, and 14,838 cases across ten categories related to household heads and items. Among the machine learning methods tested, the decision tree model outperformed the others and was identified as the most effective.

Another study by Abu Bakar et al. (2020) employed the Random Forest method to model poverty levels in Malaysia. The study used 15 factors to classify poverty: income per capita, ethnicity, state, religion, number of household members, strata, occupation, age, disability, gender, education, health, marital status, and poverty status (as the target class). Among the machine learning methods tested, Random Forest achieved the highest classification accuracy at 99.00%. Additionally, 14 poverty-related factors were aligned with the indicators from the 11th Malaysian Economic Plan and analysed using the Linear Model, Pearson Correlation, Decision Tree, and Random Forest to rank their importance. The study identified the top seven

contributing factors to poverty: income per capita, state, ethnicity, strata, religion, occupation, and education.

Voipe et al. (1996) employed the Naive Bayes method to map the potential of low-income families in Indonesia. The study aimed to identify and anticipate the poverty rate by classifying poor households. Eleven parameters were used: food, clothing, shelter, income, health, education, wealth (in rupiah), property (land), water, electricity, and the number of family members. The class labels used were "extreme poor," "very poor," and "poor." The study utilised a sample of 219 low-income families. The classification system was developed using Java and compared against results from the Weka software, achieving a classification accuracy of 93.00%. Additionally, the classification results were mapped by incorporating latitude and longitude data along with images of the houses of low-income families. The findings demonstrated that mapping with the Naive Bayes classifier could assist the government of Bantul Regency in assessing and understanding the distribution of poverty more effectively.

The studies reviewed various ML methods and their performance in classifying data related to income classes. Among these, the Random Forest method, an ensemble learning technique, demonstrated particularly promising results compared to other ML approaches. Therefore, employing Random Forest in this study using expenditure data offers a valuable new tool for the financial sector to assess financial literacy based on spending behaviour.

MATERIALS AND METHODS

The methodology for this study contains five phases: business understanding, data understanding, preparation, model development, and model evaluation. Understanding the business goals is a critical first step in data analysis. The dataset used is the 2016 HIES data obtained from the DOSM. The major steps of the study are as follows:

1. The household income and expenditure data are grouped into three income classes, namely B40, M40, and T20.
2. The data are labelled as overspending or non-overspending based on individual expenditure.
1. An ML method using the Random Forest algorithm is employed to develop the overspending classification model.
2. Overspending parameters among the income classes are ranked using the Information Gain.
3. The rules generated from the Random Forest model with the highest accuracy are analysed to gain insights from the data.

A. Business Understanding

The primary business goal of this study is to identify overspending indicators among the three income classes in Malaysia using HIES data, contributing to the classification of spending behaviour. In addition, a classification model of spending behaviour is developed using the

Random Forest method. Currently, there is a lack of models that effectively identify suboptimal financial behaviour to assist households in managing their financial planning. Furthermore, identifying the expenses that contribute most to unnecessary overspending can help individuals plan their spending more effectively.

B. Data Understanding and Preparation

The HIES 2016 data was analysed before data preprocessing and preparation. The dataset comprises household, household member, and expenditure records, containing 14,551 data rows and 163 parameters. Of these, 23 parameters capture demographic information, household income, and general expenditure, while the remaining 139 parameters detail specific household expenditure types. The 2016 HIES data were integrated with the Classification of Individual Consumption According to Purpose (COICOP). COICOP organises individual and household consumption expenditures into twelve divisions, grouping similar goods and services into consistent categories. These twelve categories are: Food and non-alcoholic beverages, Alcoholic beverages, tobacco, narcotics (ATN), Clothing, Housing, Furnishing, Health, Transportation, Communication, Recreation and culture (R&C), Education, Restaurants and hotels (R&H) and miscellaneous. The 139 specific expenditure types were aggregated and grouped under the twelve COICOP categories. Subsequently, discretisation was performed on several parameters such as age, household size, marital status, highest certification, and employment status to capture non-linear relationships, minimise the effects of differing scales and ranges, and improve the algorithm's ability to detect patterns and relationships within the data.

A new binary class parameter was added to the dataset: households that spend more than they earn are labelled as '1' (overspending), while all others are labelled as '0' (non-overspending). Several redundant or irrelevant parameters were removed from the dataset, including total income, total expenditure, level of education, relationship to the head of household, industry, occupation, and citizenship. The dataset was then segmented into three income classes: B40, M40, and T20. After this segmentation, the income parameter was discretised and replaced with income ranges, as shown in Table 1. Table 2 presents the final list of parameters used in the study, which includes all numeric variables, the income range (string), and the overspending label (binary).

TABLE 1: Income Range for Each Income Class

Income Class	Income	Income Range
B40	≤ 2500	1
	2501-3170	2
	3171-3970	3
	≥ 3971	4
M40	≤ 5880	1
	5881-7100	2
	7101-8700	3
	≥ 8701	4
T20	≤ 15040	1
	≥ 15041	2

TABLE 2. List of Parameters after Data Preparation

No	Parameter	Description
1	ID	Household identification number
2	Household Size	The number of households
3	Region	1 Peninsula, 2 Sabah and W. P. Labuan and 3 Sarawak
4	State	1 Johor, 2 Kedah, 3 Kelantan, 4 Melaka, 5 State Sembilan, 6 Pahang, 7 Penang, 8 Perak, 9 Perlis, 10 Selangor, 11 Terengganu, 12 Sabah, 13 Sarawak, 14 W.P. Kuala Lumpur, 15 W.P. Labuan and 16 W.P. Putrajaya
5	Strata	1 city; 2 rural areas
6	Type of residential	1 bungalow, 2 semi-D, 3 terraces, series or trips, city houses, 4 longhouses (Sabah & Sarawak only), 5 flats, 6 apartments, 7 condos, 8 shop/office houses, 9 rooms, 10 replacement / temporary huts and 11 other
7	Status	1 owned, 2 rented, 3 squatters owned, 4 squatters rent, 5 quarters and 6 other
8	Sex	1 male; 2 female
9	Age	1 less than 26, 2 26-60, 3 more than 60
10	Race	1 Bumiputera, 2 Chinese, 3 India, 4 Other
11	Marital Status	1 never married, 2 married, 3 widows / widows, 4 divorced, 5 separated
12	Highest Certification	1 Degree/Advanced Diploma, 2 Diploma/Certificate, 3 STPM, 4 SPM/SPMV, 5 PMR/SRP, 6 No Certificate
13	Employment	1 employer, 2 salaried worker, 3 unemployed or unpaid worker
14	Food	Percentage of spending on food
15	ATN	Percentage of spending on alcohol tobacco, narcotics (ATN)
16	Clothes	Percentage of spending on clothes
17	Housing	Percentage of spending on housing
18	Furnishing	Percentage of spending on furnishing and house maintenance
19	Health	Percentage of spending on health
20	Transportation	Percentage of spending on transportation
21	Communication	Percentage of spending on communication
22	R&C	Percentage of spending on recreation and culture
23	Education	Percentage of spending on education
24	R&H	Percentage of spending on restaurant and hotel
25	Miscellaneous	Percentage of miscellaneous spending
26	Income range	Income range within each class
27	Overspending	1 Yes, 0 No

C. Model Development with Random Forest

ML algorithms can analyse large volumes of expenditure data to identify patterns and trends in spending behaviour, offering valuable insights into financial habits and helping individuals take control of their finances (Lusardi & Mitchell, 2014). This information can be used to detect individuals who overspend or mismanage their finances. Additionally, ML algorithms can predict future financial behaviour, enabling individuals to avoid potential financial difficulties. For example, these algorithms can forecast the risk of missed loan payments or the likelihood of incurring significant unexpected expenses. Such predictive insights can support individuals in making informed financial decisions and proactively improving their economic well-being (Lin et al., 2012).

Random Forest is an ensemble ML classification method that consists of a collection of tree-structured classifiers. $\{h(x, \theta k), k = 1, ...\}$. The $\{\theta k\}$ are independently and identically distributed random vectors, and each tree casts a unit vote for the most popular class at input (Breiman, 2001). The Random Forest (RF) algorithm is a bagging ensemble classifier. It runs

efficiently and is considered to have relatively high accuracy compared to other classification algorithms (Thoplan, 2014). RF can overcome the overfitting problem because, with a large number of trees, the generalisation error converges to a limiting value under the strong law of large numbers (Breiman, 2001).

A random observations (records) sample is taken, and subsequent bootstrap samples for other trees are generated. A subset of m parameters, much smaller than the total number of parameters in the dataset, is randomly selected using the Gini score to determine the best split. The out-of-bag (OOB) prediction is obtained through a majority vote among trees for which the observation was not included in the bootstrap sample. Additionally, Random Forest can provide a ranking of parameter importance. To evaluate the importance of a parameter, Louppe et al. (2013) proposed calculating, for all trees in the forest, the average impurity decrease for all nodes where the parameter is used. The parameter with the most significant decrease in impurity is considered the most important. This can be measured using either the Mean Decrease Gini (MDG) or the Mean Decrease Accuracy (MDA) (Abu Bakar et al., 2020).

Using the notations from Louppe et al. (2013), any mean decrease impurity measure (Imp (Xm)) can be mathematically represented as shown in Equation 1:

$$\mathrm{Imp}(X_m) = \frac{1}{N_T} \sum T \sum t \in T: v(s_t) = X_m \; p(t)\Delta i(s_t, t) \tag{1}$$

From Equation (1), X_m represents the parameter of interest, N_T is the number of trees in the forest, $v(s_t)$ is the parameter at split s_t, p(t) is the proportion of records at node *t* relative to the total number of records in the dataset, and

$$\Delta i(s,t) = i\,(t) - p_L i\,(t_L) - \; p_R i\,(t_R) \tag{2}$$

p_L represents the number of records in the left child node of *t* out of the total number at node *t*. This study represents the impurity measure i (t) as the Gini index. The Gini index, *i (t)* for a node *t* is defined as in Equation (3):

$$i\,(t) = 1 - \sum_i p\,(j|t)^2 \tag{3}$$

where j = 1, 2 for this study, representing the overspending class.

D. Model Evaluation

In classification problems, accuracy, precision, and recall are commonly used to evaluate model performance. Therefore, we also employed accuracy, precision, and recall to assess our proposed approach. Knowledge of the confusion matrix is required to understand how these metrics are computed. A confusion matrix is a table that illustrates the performance of a classification task where the actual class labels are known. In our case, there are two possible

classes: whether a review contains purchase intention. Thus, a 2 × 2 confusion matrix is used, as shown in Table 3.

The number of cases correctly classified as indicating overspending or not will be placed under TP (True Positive) and TN (True Negative), respectively, while the cases incorrectly classified will be placed under FP (False Positive) and FN (False Negative), respectively.

TABLE 3: Confusion Matrix

		Classified Values	
		Positive (OS)	Negative (Not OS)
Actual Values	Positive (OS)	True-positive (TP)	False-negative (FN)
	Negative (Not OS)	False-positive (FP)	True-negative (TN)

Accuracy is a simple evaluation measure calculated as the ratio of correctly predicted cases to the total number of cases. The formula is provided in Equation (4):

$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN) \quad (4)$$

Precision is calculated as the ratio of correctly predicted positive cases to the total number of predicted positive cases. It indicates how many of the instances classified as positive are actually correct. Precision is given by Equation (5):

$$\text{Precision}=TP/(TP+FP) \quad (5)$$

Recall is the ratio of correctly predicted positive cases to the total number of actual positive cases. It indicates how many of the actual positives were correctly identified. Recall is given by Equation (6):

$$\text{Recall}=TP/(TP+FN) \quad (6)$$

RESULT AND DISCUSSION

The study yielded key findings, including the selected parameters, the best predictive models, and significant decision rules. Feature selection is crucial in data analytics as it identifies the most relevant parameters that impact the model's outcome, leading to more accurate predictions. Comparing predictive models is also essential to determine which method performs best for a given dataset and to avoid overfitting or underfitting. Finally, generating rules from the best-performing model helps interpret the results. It provides insights into the underlying patterns and relationships between parameters, which can be valuable for decision-making and problem-solving.

A. Feature Selection

Feature selection involves identifying the most important parameters or features in a dataset to predict the target variable. In this study, we employed the SelectFromModel function in the Python Sci-Kit Learn library, which uses an ML model to select the most important parameters. The function applies a mean threshold to select parameters with an importance score greater than the average of all scores. Random Forest works by building multiple decision trees on random subsets of the dataset and aggregating the predictions of each tree to produce a final output. During this process, Random Forest calculates the importance of each parameter or variable based on how much it reduces the impurity of the nodes in the decision trees (Breiman, 2001).

After building the Random Forest model, feature importance scores can be extracted and ranked. Parameters with scores higher than the mean importance are considered more relevant for predicting the class parameter, while those with scores below the mean can be removed from the dataset. This process reduces the dataset's dimensionality, making it easier to work with and potentially improving the model's performance. The selected parameters for each dataset are compiled and used to test classification models. Table 4 presents the important parameters identified for the B40, M40, and T20 income class datasets. Descriptions of these parameters can also be found in Table 4.

TABLE 4. Selected Overspending Parameters for B40, M40, and T20 Income Classes

No Parameters	B40	M40	T20
1	Housing	R&H	R&H
2	Food	Miscellaneous	Miscellaneous
3	R&H	Furnishing	House
4	Furnishing	Food	R&C
5	Transportation	House	Furnishing
6	Clothes	Transportation	Transportation
7	Miscellaneous	R&C	ATN
8	Health	Clothes	Clothes
9	R&C	Health	Food
10	ATN	ATN	Health

B. Classification Models

The selected parameters from the Random Forest algorithm were used to build predictive models, including Decision Tree, Random Forest, Support Vector Machine, k-Nearest Neighbours (kNN), Naive Bayes, and Gradient Boosted Classifier. The overall metric scores of each model are shown in Table 5. The results indicate that Random Forest outperformed the other methods across all evaluation measures, with the exception of recall, where it was slightly lower than kNN. It is worth noting that Random Forest is generally less prone to overfitting than many other ML methods. However, if the individual decision trees within the Random Forest are too complex or if the ensemble contains too many trees, it may be necessary to control the growth of subtrees.

TABLE 5. Metric Scores of Overspending Classification Models, Highlighting Random Forest as the Best Performer

	Machine Learning Methods					
Evaluation Metrics	DT	RF	SVM	kNN	NB	GB
Accuracy	0.893	0.921	0.712	0.895	0.688	0.886
Precision	0.872	0.914	0.724	0.866	0.763	0.877
Recall	0.927	0.937	0.750	0.955	0.573	0.906
F1 Score	0.899	0.925	0.731	0.906	0.645	0.891
ROC AUC	0.891	0.919	0.703	0.894	0.686	0.885

Footnote: DT=Decision Tree, RF=Random Forest, SVM=Support Vector Machine, kNN=k-nearest neighbour, NB=Naïve Bayes, GB= Gradient Boosted

C. Rules Generation

To generate rules from the Random Forest model, trees with similar top-level nodes—those with the highest importance were identified. This approach ensures that the generated rules are based on the most significant parameters in the dataset. Next, branches containing valuable rules and low Gini indices were extracted from the selected trees. The Gini index measures the degree of impurity or randomness in a decision tree or Random Forest model. A lower Gini index indicates a better-performing model with more accurate predictions and less randomness in its decision-making process.

Rules generation for the B40 dataset involves identifying the first three nodes of the decision tree that correspond to the top three most important parameters. In this case, the parameters with the highest importance are income range, housing expenses, and transportation expenses. The first three node splits of the decision tree are associated with these parameters. Figure 1 illustrates an example of one of the decision trees generated by the Random Forest model, and the rules extracted from this tree are listed in Table 6. For example, Rule 1 suggests that a specific combination of spending percentages on Housing, Transportation, Furnishing, and Miscellaneous contributes to overspending. The income range and state are contextual details associated with these rules. The rules derived for the B40 income class reveal the ranking of overspending items, based on the maximum percentage of overspending extracted from the rules. These are: Housing (18.9%), Food (16.1%), R&H (11.1%), Furnishing (9.67%), Transportation (8.9%), Clothes (4.57%), Miscellaneous (3.94%), Health (1.27%) and R&C (0.47%). A comparative analysis with the baseline models will be discussed in Part D in this section.

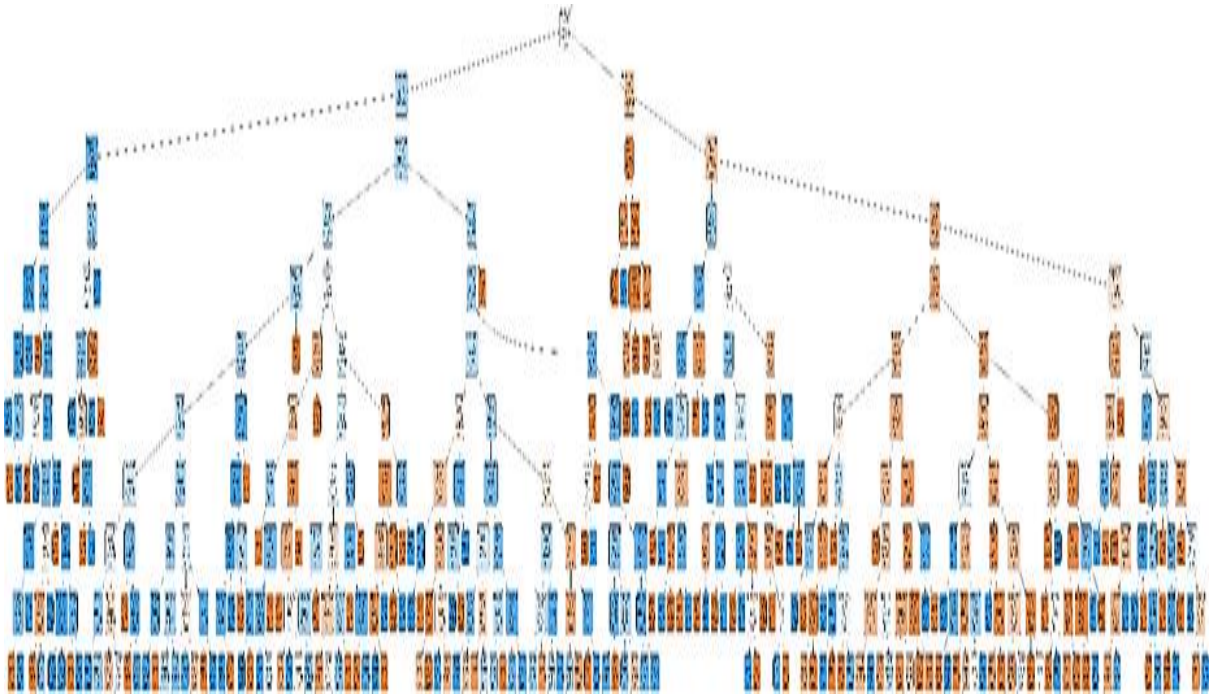

FIGURE 1. Illustration of the Selected Decision Tree (Tree 66) from the Random Forest Model for B40 Overspending Classification for Rule Extraction

TABLE 6. Rules for Overspending Generated from the B40 Classification Model

No	Rules Generated from Selected Tree(s)
1	IF Income Range ≤ 1.5 AND Housing > 18.9 AND Transportation> 8.92 AND State > 11.5 AND Furnishing> 9.67 AND Miscellaneous 3.94
2	IF Income Range ≤ 1.5 AND Transportation= 8.93-9.25 AND Housing > 18.09 AND State<=11.5 AND R&H<=15.02
3	IF Income Range ≤ 1.5 AND Housing > 13.22 AND Transportation< 8.92 AND Furnishing> 5.84 AND State =11.6-13.5 AND Miscellaneous>2.31
4	IF Income Range ≤ 1.5 AND Housing ≤ 13.22 AND Transportation≤ 16.5 AND Clothing ≤ 4.57 AND R&H = 11.11-18.34 AND Health > 0.84
5	IF Income Range > 2.5 AND R&C > 0.01 AND Housing >12.82 AND Furnishing>6.07 AND Transportation>9.47 AND Food > 16.14 AND Health > 1.27 AND Miscellaneous> 6.84
6	IF Income Range > 2.5 AND R&C > 0.47 AND Housing >12.82 AND Furnishing>6.07 AND Transportation>9.47 AND Food >16.14 AND State <=11.5 AND R&H >5.44

Rules generation for the M40 dataset is based on the trees identified as having the two most important parameters, namely housing and transportation, as the first two nodes. Five trees that satisfied the specification were identified (Tree 17, 28, 49, 68 and 70) and used to generate rules to predict overspending behaviour in the M40 dataset. For example, rule 1 in Table 7 interprets that the combination of specific spending percentages of Housing, Transportation, Food, Clothing, Furnishing and Miscellaneous contributes to overspending. The rules of the M40 class indicate the rank of overspending items are R&H (23.62%), Miscellaneous (19.23%), Furnishing (18.72%), Food (16.78%), Housing (15.71%), Transportation (6.99%), R&C (5.32%), Clothes (4.09%), Health (0.95%). The values in the bracket are the maximum % of

overspending extracted from the rules. The parameter state is the information related to the rules. Part D of this section will describe the comparative analysis with the baselines.

TABLE 7. Rules for Overspending Generated from the M40 Classification Model

No	Rules Generated from Selected Tree(s)
1	IF Housing > 13.73 AND Transportation> 4.09 AND Food > 16.78 AND State > 4.5 AND Clothing = 2.77 AND Furnishing= 10.73-11.47
2	IF Housing > 13.73 AND Transportation> 6.99 AND Food > 16.78 AND State = 1.5-4.5 AND Furnishing>18.72
3	IF Housing > 11.8 AND Transportation> 4.29 AND State > 6.5 AND R&H > 20.74 AND Furnishing> 7.45 AND Health > 0.95 AND Clothing > 4.09
4	IF Housing >13.73 AND R&H > 23.62 AND Transportation> 10.11 AND Furnishing> 1.61 AND Clothing = 2.37
5	IF Housing > 15.71 AND Transportation> 4.08 AND Clothing>3.03 AND Miscellaneous>19.23 AND R&C > 2.10
6	IF Housing > 22.79 AND Transportation> 4.08 AND State ≤ 4.5 AND Health > 0.33 AND Miscellaneous> 9.63 AND R&C = 5.32-5.56

The rule generation for the T20 dataset was carried out by identifying the trees in which the two most important parameters, Housing and R&H, appear as the first two nodes. Five trees met this criterion: Tree 3, 9, 25, 44, and 46. For example, in Table 8, Rule 1 indicates that a specific combination of spending percentages on Housing, Transportation, R&H, and Miscellaneous contributes to overspending. The rules derived for the T20 class reveal the ranking of overspending items as follows: R&H (28.78%), Miscellaneous (20.02%), Housing (10.76%), R&C (10.67%), Furnishing (9.89%), and Transportation (5.22%). The values in parentheses represent the maximum percentage of overspending extracted from the rules. A comparative analysis with the baseline results will be discussed in Part D in this section.

TABLE 8. Rules for Overspending Generated from the T20 Classification Model

No	Rules Generated from Selected Tree(s)
1	IF Housing > 10.76 AND R&H > 15.51 AND Miscellaneous > 20.02 AND Transportation ≤ 5.22
2	IF Housing > 10.23 AND R&H > 28.78 AND Transportation≤ 4.9 AND Furnishing > 9.89
3	IF Housing > 10.23 AND R&H > 15.5 AND Furnishing = 9.92-10.02
4	IF Housing > 10.23 AND R&H > 15.5 AND Furnishing > 9.89 AND R&C > 10.67
5	IF Housing > 10.23 AND R&H >15.5 AND Transportation ≤ 4.9 AND Furnishing > 8.36

The Random Forest model has demonstrated higher accuracy than other methods in predicting overspending. The important parameters extracted from the B40, M40, and T20 datasets share some commonalities, but each group also exhibits unique characteristics. In the B40 dataset, the income range is one of the selected parameters, highlighting its contribution to the risk of overspending within this income group. This finding is supported by a study conducted by Othman et al. (2020), which reported that 75% of individuals in the B40 group with incomes below RM2,768 fall into the overspending category. Low-income households are more likely to overspend because their monthly income is often insufficient to cover their living expenses.

For the M40 rules, in addition to overspending items, the state parameter also plays an important role in the classification rules. This suggests that M40 households from different states exhibit varying overspending patterns. In the T20 dataset, the R&H parameter (i.e., Restaurants and Hotels) stands out, indicating that the T20 group has greater financial freedom

compared to the B40 and M40 groups, allowing them to spend more on luxury dining and hotel accommodations.

The rules generated for the B40 group indicate a tendency to overspend on furnishing, restaurants and hotels (R&H), and miscellaneous items. Additionally, the state of residence contributes to overspending among B40 households. The decision tree often separates Peninsular Malaysia (excluding the federal territories) from Sabah, Sarawak, and the federal territories at a node. The rules suggest that low-income households in Sabah and Sarawak are more likely to overspend, despite having lower transportation expenses. Overall, this analysis provides valuable insights into the factors contributing to overspending among B40 households.

The rules for the M40 dataset are particularly interesting, as they highlight overspending factors such as furnishing, health, clothing, food, housing, transportation, and restaurants and hotels (R&H). This model also identifies the state of residence as an important factor influencing spending behaviour. Meanwhile, the rules generated for the T20 dataset reveal that spending on furnishing is a key contributor to overspending within this group. However, beyond this, the model offers limited valuable insights, likely due to class imbalance in the dataset. High-income households are generally less prone to overspending than low-income households, resulting in fewer overspending in the data. Consequently, the model lacks sufficient examples to generate meaningful rules and insights for the T20 group.

D. Comparative Analysis of Overspending Items in the Random Forest Model

The effectiveness of the Random Forest overspending model, particularly its identification of overspent items, is assessed against household expenditure data from 2019 and 2022. A general analysis of overspending items identified by the model is conducted by cross-referencing them with the percentage of spending reported in the 2016, 2019, and 2022 Household Expenditure Survey Reports. The objective is to evaluate the model's ability to predict future spending behaviour. Tables 9 to 11 present a comparison between the overspending parameter rankings generated by the proposed Random Forest model (refer to Table 4) and the expenditure rankings based on the highest percentage values from the DOSM reports for 2016, 2019, and 2022. Additionally, our model provides predicted overspending percentages for each item, shown in the two rightmost columns. These insights highlight the model's potential in forecasting overspending behaviour across different income classes.
The state and income parameters are excluded in this section, as they do not represent spending items. The Random Forest parameters are ranked in descending order based on the percentage of spending extracted from the selected rules, ensuring consistency with the baseline presentations. Although the overspending percentages generated by the Random Forest model are not directly comparable to the expenditure percentages reported by DOSM due to the use of data analytics methods, they are included in the tables to enable comparison of the parameter importance rankings for both spending and overspending items.

In Table 9, the overspending items identified by the proposed Random Forest (RF) model show that the B40 group overspent the most on housing, food, and restaurants & hotels (R&H). Additionally, furnishings and transportation are also among the top five ranked items. The

percentages of overspending in the RF model were derived from the rules presented in Table 6. The ranking of overspending items based on the 2016 data aligns with the top five expense items reported in the 2019 and 2022 DOSM reports. Notably, R&H spending surpasses transportation in the 2019 and 2022 reports, indicating a shift in lifestyle among households over the six years. The proposed model effectively extracts meaningful patterns from the 2016 data that align with spending behaviours observed in later years. In other words, the overspending features from the 2016 data can predict household expenditure trends over the following six years.

TABLE 9. Comparison of Parameter Importance from the RF Model with DOSM Data (2016, 2019, and 2022) for the B40 Dataset

	Spending Items of Household Expenditure DOSM Reports (2016, 2019, 2022)						Over-spending Items RF Model 2016	
Rank	2016	(%)	2019	(%)	2022	(%)	2016	(%)
1	Food	25.5	Food	25.6	House	25.6	House	18.9
2	House	24.7	House	24.2	Food	24.5	Food	16.1
3	Transportation	11.8	R&H	12.6	R&H	12.9	R&H	11.1
4	R&H	11.8	Transportation	10.8	Transportation	10.0	Furnishing	9.67
5	Miscellaneous	6.5	Miscellaneous	6.8	Communication	5.6	Transportation	8.9
6	Communication	4.0	Communication	4.2	Miscellaneous	5.4	Clothes	4.57
7	Clothes	3.6	Furnishing	3.6	Furnishing	4.1	Miscellaneous	3.94
8	R&C	3.5	Clothes	3.5	Clothes	2.9	Health	1.27
9	Furnishing	3.2	R&C	3.5	Health	2.8	R&C	0.47
10	ATN	2.5	ATN	2.3	ATN	2.1	-	-
11	Health	1.8	Health	2.0	Service	2.0	-	-
12	Education	1.0	Education	0.9	Education	0.8	-	-

In Table 10, for the M40 income group, the overspending Random Forest model identifies the top five overspending items as R&H, Miscellaneous, Furnishing, Food, and Housing. These items were also reported as among the most significant expenditures for M40 households in the 2016, 2019, and 2022 DOSM Household Expenditure Reports. The percentages of overspending for the Random Forest model were obtained from the rules in Table 7. The model highlights that M40 households tend to overspend most on R&H, Miscellaneous, and Furnishing, indicating that this group is more flexible in allocating spending to these areas.

The findings suggest that the overspending model proposed in this study aligns well with actual expenditure patterns, as the items identified in the 2019 and 2022 reports are consistent with the overspending predictions. It indicates that the model effectively captures overspending behaviour within the M40 group, which differs significantly from the B40 overspending model. The difference may be attributed to the greater disposable income available to M40 households.

TABLE 10. Comparisons of Parameter Importance from the RF Model with DOSM Data (2016, 2019, 2022) for the M40 Dataset

	Spending Items of Household Expenditure DOSM Reports (2016, 2019, 2022)						Over-spending Items RF Model 2016	
Rank	2016	(%)	2019	(%)	2022	(%)	2016	(%)
1	House	22.8	House	22.8	House	22.25	R&H	23.6
2	Food	19.0	Food	18.0	Food	17.3	Miscellaneous	19.2
3	R&H	13.9	R&H	14.5	R&H	16.6	Furnishing	18.7
4	Transportation	13.8	Transportation	13.5	Transportation	10.95	Food	16.7
5	Miscellaneous	7.6	Miscellaneous	8.0	Communication	6.85	House	15.7
6	Communication	5.2	Communication	5.2	Miscellaneous	6.0	Transportation	7.0
7	R&C	4.7	R&C	4.7	Furnishing	4.65	R&C	5.3
8	Furnishing	3.9	Furnishing	4.2	Service	3.75	Clothes	4.1
9	Clothes	3.4	Clothes	3.3	Clothes	2.9	Health	1.0
10	ATN	2.6	ATN	2.4	Health	2.75	ATN	-
11	Health	1.9	Health	2.0	R&C	2.7	-	-
12	Education	1.3	Education	1.4	ATN	1.95	-	-

Table 11 shows that the Random Forest overspending model for the T20 group identifies the top five items as R&H, Miscellaneous, Housing, R&C, and Furnishing. Except for R&C and Furnishing, these items were also reported in the 2016, 2019, and 2022 DOSM Household Expenditure Reports as among the most significant expenses for T20 households. The percentage of overspending in the Random Forest model was derived from the rules in Table 8. Including R&C and Furnishing as overspending items suggests that this group has greater financial freedom in these areas, which may contribute to financial literacy concerns. Similar to the M40 group, the proposed overspending model does not significantly predict the actual expenses of the T20 group in 2019 and 2022, as reflected in the discrepancies in item rankings.

TABLE 11. Comparisons of Parameter Importance from the RF Model with DOSM Data (2016, 2019, 2022) for T20 Dataset

	Spending Items of Household Expenditure DOSM Reports (2016, 2019, 2022)						Over-spending Items RF Model 2016	
Rank	2016	(%)	2019	(%)	2022	(%)	2016	(%)
1	House	24.4	House	22.2	House	23.3	R&H	28.78
2	Transportation	15.4	Transportation	15.5	R&H	17	Miscellaneous	20.02
3	R&H	13.8	R&H	13.8	Transportation	12.5	House	10.76
4	Food	12.2	Food	12.6	Food	11.2	R&C	10.67
5	Miscellaneous	8.8	Miscellaneous	8.9	Communication	6.6	Furnishing	9.89
6	R&C	6.2	R&C	6.6	Miscellaneous	6.6	Transportation	5.22
7	Communication	5.4	Communication	5.3	Service	5.3	ATN	-
8	Furnishing	5.1	Furnishig	5.2	Furnishing	5.2	Clothes	-
9	Clothes	3.3	Clothes	3.5	R&C	4.1	Food	-
10	ATN	2	Health	2.4	Health	2.7	-	-
11	Health	1.9	ATN	2.1	Clothes	2.4	-	-
12	Education	1.5	Education	1.9	Education	1.6	-	-

The DOSM reported that the T20 and M40 groups have greater flexibility in determining their spending patterns, unlike the B40 group, which is constrained to allocate expenses primarily for basic needs due to limited income. The findings of this study support these statements and

further highlight additional socio-economic characteristics that contribute to overspending (DOSM, 2020). Analysing the Random Forest models developed in this study alongside the DOSM statistical reports reveals comparable rankings of expenditure items. However, the Random Forest overspending model identifies certain items as significantly more overspent than others. Notably, the overspending items identified for the B40 group closely align with the 2019 and 2022 DOSM reports. It suggests that the machine learning model trained on 2016 data has predictive capability for forecasting future spending behaviour reflected in the 2019 and 2022 household expenditure data.

In addition to the general comparative analysis, two significant differences between our Random Forest model and the statistical information provided by DOSM are noteworthy. First, the ML model through Random Forest derives actionable insights from data to predict future or previously unknown events. Second, while statistical analysis primarily evaluates the validity and significance of existing information, predictive analytics enables the analysis of large datasets to build models that forecast future outcomes. Despite minor differences in parameter rankings between the two approaches, the Random Forest model offers the advantage of predicting overspending behaviour through the knowledge extracted in the form of decision rules. Although both methods identify similar top parameters such as Housing, R&H, Transportation, and Food, the ML approach provides deeper insights by capturing the combinations of these parameters that characterise overspending behaviours across different household income classes.

E. The Insights

The proposed Random Forest model generates rules that provide valuable insights for financial educators in designing effective interventions to promote financial literacy. While the model shows promise, it can be further enhanced to achieve more accurate classifications. Data preparation, in particular, can be improved in two key ways. First, expenditure threshold values for each item should be defined to distinguish between overspending and non-overspending within each income class. For example, determining the maximum acceptable spending on housing for the B40, M40, and T20 groups would enable more precise classification. Second, developing a more balanced dataset would help create a model that fairly represents all income classes. Furthermore, incorporating the 2019 and 2022 household income and expenditure data could enhance the model's predictive power, allowing for more accurate forecasting of future household expenses and spending behaviour.

Additionally, low-income individuals can improve their spending behaviour and reduce the risk of overspending by creating a budget, prioritising essential expenses, minimising debt, and seeking guidance from a financial advisor or therapist. By taking a proactive approach to financial well-being, individuals can manage overspending, strengthen their financial position, and work toward long-term stability. In conclusion, overspending can significantly affect the financial well-being of low-income individuals. To mitigate this impact, coordinated efforts from policymakers, financial institutions, and individuals are essential to promote financial literacy and provide access to relevant resources and support. With the right tools and strategies, individuals can overcome overspending, build financial resilience, and achieve financial goals.

CONCLUSION

This paper presents an experimental study exploring the application of machine learning (ML) in leveraging socio-economic data. We employed the Random Forest method to build a classification model for overspending among household income classes using Malaysia's 2016 household income and expenditure data. The Random Forest model accurately predicts overspending behaviour, and the generated rules offer valuable insights into overspending patterns. Notably, the generalisation of the model trained on 2016 data successfully predicted household spending behaviours in 2019 and 2022. The study demonstrates the potential of ML in financial research and underscores the importance of balancing model interpretability with predictive performance. One key challenge in analysing financial behaviour is its complex and multifaceted nature, which cannot be fully explained by overspending alone. Additional factors such as budgeting, saving, investing, and borrowing must be considered to construct a more comprehensive model of financial behaviour. Incorporating such data into the model would enhance its accuracy and reliability if such data becomes available. Integrating information on existing savings, investments, and loans would allow the development of a more precise classification of financial behaviour. Ultimately, this model could serve as a valuable tool to identify individuals in need of financial education or support, promoting better financial literacy and healthier financial practices.

ACKNOWLEDGEMENT

This work partially contributes to the research funded by LRGS/1/2020/UKM/01/5/2 grant under the Ministry of Higher Education, Malaysia.

REFERENCES

A. A., Bakar, Hashim R. S., Aziz J., Langmeri L. H., and Yusof S. "Kelas Pendapatan Dan Risiko Beban Kewangan Isi Rumah di Malaysia." 1st Ed. Penerbit UKM, 2021.

Abu Bakar, Azuraliza, Rusnita Hamdan, and Nor Samsiah Sani. "Ensemble Learning for Multidimensional Poverty Classification." Sains Malaysiana 49, no. 2 (February 28, 2020): 447–59. https://doi.org/10.17576/jsm-2020-4902-24.

Breiman, Leo. Machine Learning 45, no. 1 (2001): 5–32. https://doi.org/10.1023/a:1010933404324.

Creamer, Germán G. and Freund, Yoav. Predicting Performance and Quantifying Corporate Governance Risk for Latin American ADRs and Banks. Financial Engineering And Applications, MIT, Cambridge, 2004. Available at SSRN: https://ssrn.com/abstract=743209.

DOSM. "Laporan Survei Pendapatan Isi Rumah Dan Perbelanjaan 2019." DOSM, 2020. https://www.dosm.gov.my/v1/uploads/files/1_Articles_By_Themes/Prices/HIES/HIS-Report/HIS-Malaysia-.pdf.

Huang, R., M. Samy, H. Tawfik, and A.K. Nagar. Application of Support Vector Machines in Financial Literacy Modelling. 2008. Second UKSIM European Symposium on Computer Modeling and Simulation, September 2008, 311–16. https://doi.org/10.1109/ems.2008.84.

Huang, R., M. Samy, H. Tawfik, and A.K. Nagar. Application of Support Vector Machines in Financial Literacy Modelling. 2008. Second UKSIM European Symposium on Computer Modeling and Simulation, September 2008, 311–16. https://doi.org/10.1109/ems.2008.84.

Lariviere, B., and D. Vandenpoel. Predicting Customer Retention and Profitability by Using Random Forests and Regression Forests Techniques. 2005. Expert Systems with Applications 29, no. 2: 472–84. https://doi.org/10.1016/j.eswa.2005.04.043.

Levantesi, Susanna, and Giulia Zacchia. 2021. Machine Learning and Financial Literacy: An Exploration of Factors Influencing Financial Knowledge in Italy. *Journal of Risk and Financial Management* 14(3), 120. https://doi.org/10.3390/jrfm14030120.

Lin, Wei-Yang, Ya-Han Hu, and Chih-Fong Tsai. 2012. Machine Learning in Financial Crisis Prediction: A Survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C. Applications and Reviews 42(4): 421–36. https://doi.org/10.1109/tsmcc.2011.2170420.

Louppe, Gilles, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. 2013. Understanding Variable Importances in Forests of Randomised Trees. *Advances in Neural Information Processing Systems* 26.

Lusardi, Annamaria, and Olivia S. Mitchell. 2014. The Economic Importance of Financial Literacy: Theory and Evidence. *Journal of Economic Literature* 52(1): 5–44. https://doi.org/10.1257/jel.52.1.5.

Othman, Zulaiha Ali, Azuraliza Abu, Nor Samsiah, and Jamaludin Sallim. 2020. Household Overspending Model amongst B40, M40 and T20 Using Classification Algorithm. *International Journal of Advanced Computer Science and Applications* 11(7). 392-399. https://doi.org/10.14569/ijacsa.2020.0110751.

Sood, Meenakshi, and Puneet Bhushan. 2017. Efficacy of Artificial Neural Network for Financial Literacy Prediction. *International Journal of Advanced Research in IT and Engineering* 6(2): 1-8.

Thoplan, Ruben. Random Forests for Poverty Classification. 2014. *International Journal of Sciences: Basic and Applied Research (IJSBAR)* 17(2): 252-59. https://www.gssrr.org/index.php/JournalOfBasicAndApplied/article/view/2574.

Voipe, Ronald, Haiyang Chen, and Joseph Pavlicko. Personal Investment Literacy Among College Students: A Survey. Financial Practice and Education 6 (1996).

Wei-Yang Lin, Ya-Han Hu, and Chih-Fong Tsai. 2012. Machine Learning in Financial Crisis Prediction: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42(4): 421–36. https://doi.org/10.1109/tsmcc.2011.2170420.

Zainal, Nor Rashidah, Gurmit Kaur, Nor ‘Aisah Ahmad, and Jamaliah Mhd. Khalili. 2012. Housing Conditions and Quality of Life of the Urban Poor in Malaysia. *Procedia - Social and Behavioral Sciences 50* (2012): 827–38. https://doi.org/10.1016/j.sbspro.2012.08.085.

Zou, Zhi Bin, Hong Peng, and Lin Kai Luo. “The Application of Random Forest in Finance.” Applied Mechanics and Materials 740 (March 2015): 947–51. https://doi.org/10.4028/www.scientific.net/amm.740.947.