Household Overspending Model Amongst B40, M40 and T20 using Classification Algorithm

Zulaiha Ali Othman[1], Azuraliza Abu Bakar[2]
Nor Samsiah Sani [3]
Center for Artificial Intelligence Technology
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

Jamaludin Sallim[4]
Faculty of Computing
Universiti Malaysia Pahang
Gambang, Malaysia

***Abstract*—The family economy is a critical indicator of the well-being of a family institution. It can be seen by the total income and how well the household finances is managed. In Malaysia, the household income level is categorized as B40, M40 and T20. These categories can also indicate the poverty level of the household. Overspending is a phenomenon where the monthly expenses are more than the household's total income, which affects economic wellbeing. Finding important factors that affect the spending patterns among the household can reveal the causes of overspending. It will assist the government in mitigating such problems. Availability of 4 million household expenditure records obtained from the survey conducted in 2016 by the Department of Statistics Malaysia eases the aim of this study to develop a household overspending model by using machine learning. The model is developed using 12 household demographic attributes with 14451 household records. The attributes are the number of households, area, state, strata, race, highest certificate, marital status, gender, housing, income, total expenditure, and category as attributes class. The model development employs five machine learning algorithms namely decision tree, Naïve Bayes, Neural network, Support Vector Machines, Nearest Neighbour. The results show that the decision tree through J48 algorithm has produced the easiest rule to be interpreted. The model shows four attributes which were income, state, races and number of households that highly influence the overspending problem. Based on the research finding, it can be concluded that these attributes are essential for improving the indicator measure for Malaysian Family Wellbeing Index in the aspect of overspending.**

***Keywords*—Overspending; classification; poverty; household**

I. INTRODUCTION

Malaysian Family wellbeing index consisted of 8 indicators, with the economy being one of it [1]. The wellness of the economy can be measured based on two indicators; income and expense. When a person spends more than his/her income, the phenomenon is called overspending. Overspending is a continuous issue for ages [2][3], which resulted from verities of factors making the issue to be rapidly increased yearly [4]. Overspending may cause by the basic need of a family. However, it also due to lifestyle with the attitude of not being able to self-measure and being critical on the lifestyle. Undeniably, overspending can occur due to an unexpected event that might occur once in a blue month. However, it was not supposed to be happening on a monthly basis.

Overspending is expected to be experienced by most millennials due to the lack of knowledge on how to spend money wisely. With various facilities such as multiple bank accounts, insurance and saving plan causes millennials to easily trapped in overspending. Moreover, with high-cost university fees, youngsters nowadays are facing huge debt even before securing a job. Additionally, unlimited access to online shopping also contributes to this matter. Other than that, using delivery food services causes food expenses to be increased. Bankruptcy is an even more serious consequence of overspending. Forty percent of households in the USA faces with overspending since 1990. The number of non-business bankruptcies in the United States reported being increasing [5]. Malaysian also shows a huge number on this matter. The number is increasing wherein 2018, a total of 303,415 bankruptcies was reported [6].

The phenomenon of over-spending needs to be addressed, as it can lead to social problems due to financial constraints which can result in theft, unauthorized money lending, and long-term personal loans. Limited financial resources, growing of needs and the rising cost of living are challenges that young people face in order to balance their current and future needs. With a variety of financial facilities, especially credit cards, banking, investment, loan and e-wallet, it requires consumers to have the financial knowledge to use the facilities. A variety of basic needs such as emergency care, child education, credit and risk management (insurance/takaful), retirement planning and estate planning with limited resources are a challenge for today's financial management which increases the cost of living as well.

A preliminary study on consumer financial 2016 survey data shows that 8.44% of Malaysians fall into the category of over-spending. However, studies on overspending issues is still limited in Malaysia. Several smart financial management have been taught and used but many of them were focusing on financial management and less focus on the spending style [7]. Various analysis has been conducted on the data, most of the researchers focus on the type of spending [7] and still limited study focus on overspending especially using an artificial intelligent data analytic method to overcome meaningful knowledge. Research on overspending lifestyle has been conducted and found out three main significant factors that influencing spending which was food, housing and

transportation. However, the study conducted using regression and not based on the overspending context.

Thus, this study was conducted in order to develop a household over-spending model amongst Malaysians' B40, M40 and T20 using classification techniques. Thus, it will show the overspending pattern among B40, M40 and T20. The model was developed based on the demographic information only as given by the Department of Statistic Malaysia. The contribution of this work beneficial to see the serious factor of overspending issues in a family context. The finding also can be used in finding possible indicators for the multidimensional index in Malaysia.

The rest of this paper is organized as follows: Section 2 presents related work on the overspending data classification. Section 3 presents the material and method used for data classification. Section 4 reports the experimental results and Section 5, concludes the finding.

II. Related Work

Household income management is needed for a wise spending regime. Due to budget constraints, households need to plan and prioritize basic necessities. Basic necessities are defined as daily needs which includes, food, housing, transportation, healthcare and clothing [8]. It was found that low-income households spend most of their income on basic necessities rather than on unnecessary expenses. However, they are still facing overspending.

According to Rashid et al. (2018), based on all income groups and strata, there are three types of household expenditure. First, food and non-alcoholic beverages. Second, housing, water, electricity, gas and other fuels. Third, transportation. These three groups can be classified as basic necessities. Analysis shows that group B40 spends almost two-thirds of their total expenditure on non-alcoholic food and beverages, housing, water, electricity, gas, fuel, and transportation.

In 2019, [9] reported that Malaysian household spends 69.1% in four main groups which include, housing, water, electricity, gas, fuels, non-alcoholic, restaurant and hotel. On the statistical report, the highest contributors to overall consumption were for housing, water, electricity, gas and fuels (24.0%), followed by food and non-alcoholic beverages (18.0%), transportation (13.7%), restaurant and hotels (13.4%).

Rashid et al. (2018) used three regression approach to analyse the relationship between total spending and basic needs among three income group households. Result shows that spending on basic needs has a significant relationship with the total expenditure between-group income. The basic needs of food, transportation and housing showed a significant relationship with total expenditure. In other words, by increasing spending on basic needs will increase household spending. However, the researches still use basic statistical analysis.

Artificial intelligent and data analytic has known as a popular approach where discovering accurate and meaningful knowledge in various domain utilising huge pass data made possible. Its offers various task such as classification, clustering, prediction, diagnostic, and deviation detection [10]. Where, the selection of the methods is depending what kind of business problem and type of data available. The aims are to identify the best model that gives the highest classification accuracy. Besides measure of accuracy, other measure such as mean absolute error, Root Mean Squared Error (RMSE), F-measure, Precision, Recall, the Kappa statistic, ROC and computation time are also considered in evaluating the performance of the model.

Classification usually used for predicting or discovering new knowledge in a form of rules, tree or function [11][12]. There are various algorithms that fall under classification technique such as J48, Naïve Bayes, Neural network, Support Vector Machines and Nearest Neighbour [13]. J48 is an enhancement of the C4.5 decision tree algorithm which functions by creating decision tree that based on data attributes. This algorithm identifies the attribute that discriminates instances most clearly which. The quality of rule, tree or function created from this algorithm can be determined by the accuracy of the model [14]. J48 had been proven to having highest accuracy compared to other algorithm. In analysing poverty level in Indonesia, [15] had done study using J48. Another study done using random forest algorithm in measuring poverty in urban area which provide more directional and timely decision-making assistance for the resource allocation and renewal planning of poor communities [16].

Neural Network is a mathematical model or computational model based on emulation of a biological neural system. There are several neural network algorithms such as ANN, CNN and kNN. The output value of the neuron is usually a non-linear transformation of the sum of stimuli. In more advanced models, the non-linear transformation is adapted by some continuous functions. NN is very popular for prediction with few attributes such as stock market prediction [17], weather forecasting [18] and customer churn [19]. NN was used by [20] in mapping out the poverty in Mexico. Where, CNN was used in predicting poverty mapping for urban areas using imagery for Digital Globe or Planet.

Another algorithm that usually used in study is Bayesian. This algorithm involves statistical methods that assign probabilities or distributions to events or parameters based on experience or best guesses before experimentation and data collection. A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be the "independent feature model" [21]. Bayesian has shown as an accurate model for various problem [22][23]. Naïve Bayes was used by [24] in mapping out the potentially poor family in Indonesia to planning the right method in preventing such occurrence towards the family.

Another classification technique that is popular is Support Vector Machine (SVM), which can be employed for both classification and regression purposes. SVM works and completes the analysis through a series of binary assessments on the data. SVM has shown as a good algorithm in the various domain is very popular particularly in image processing.

Thus, this study was conducted in order to develop a household over-spending model amongst Malaysians' B40, M40 and T20 using classification techniques that shows overspending pattern. The model was developed based on the demographic information as given by the Department of Statistic Malaysia (DOSM). The contribution of this work beneficial to see the serious factor of overspending issues in a family context. The finding also can be used in finding possible indicators for the multidimensional index in Malaysia.

III. Material and Methods

The study follows the standard data mining step of three phases which are (1) defining business goal, (2) data collection, (3) data preprocessing and preparation, and (4) development of model [25].

A. Defniing Business Goal

In this study, we define the business goals as to identify the patterns of overspending among various household income classes (B40, M40, and T20). Factors that contributes to the goal will be identified through features selection from source data set.

B. Data Collection

Survey was conducted in 2016 by the Department of Statistics Malaysia (DOSM), with a total of 4 million household expenditure records were obtained. However, only 20 percent of the data were used in this study due to constraints in obtaining all of the data from the DOSM. Data obtained consisted of 12 attributes on demographics data selected. They were number of households, area, state, race, highest certificate, marital status, gender, housing income, total expenditure and category which were described in detail in [26].

C. Preprocessing and Preparation

Phases involve in data preparation were namely attribute selection and class label determination. There were several phases involved in this study which discussed as follows.

The first phase was data preparation. This phase was done by identifying as much as attributes and records can be collected. Then the data cleaning process was done which include generating new attributes of poverty and the overspending category. In his phase, the cleaning process which include replacing incomplete and incorrect data with null was done.

The second phase was descriptive analysis using SQL language. The analysis was done towards the number the percentages and distribution in each state and the total spends for each category. The analysis was then used in overcoming basic knowledge on the overspending pattern amongst B40, M40 and T40.

The third phase was pre-processing by discretising data into the nominal form of attributes. Then, determining the best modelling followed by interpreting the knowledge was conducted.

1) Income class level

a) Generate income class: The income class was generated by referring to the amount of income and state set by the Malaysian government [26]. The algorithm was translated into the rules in Table I.

TABLE I. Rules for Income Class

IF State = ‘Melaka’ OR ‘P.Pinang’ OR ‘Johor’ AND IF income < 4768.92 THEN Category = B40 ELSE IF 4768.92 =<income <9380.15 THEN Category=M40 ELSE Category = T20
IF Negeri = ‘Perlis’ OR ‘Perak’ OR ‘Pahang’ AND IF income < 3461.75 THEN Category = B40 ELSE IF 3461.75=<income <6814.58 THEN Category=M40 ELSE Category = T20
IF State = ‘W.P Putrajaya’ AND IF income < 6814.58 THEN Category = B40 ELSE IF 6814.58=<income <15170.35 THEN Category=M40 ELSE Category = T20
IF State = ‘Sabah’ AND IF income < 3180.85 THEN Category = B40 ELSE IF 3180.85=<income <7622.05 THEN Category=M40 ELSE Category = T20
IF State = ‘W.P Labuan’ AND IF income < 4768.92 THEN Category = B40 ELSE IF 4768.92=<income <12435.15 THEN Category=M40 ELSE Category = T20
IF State = ‘W.P KL’ AND IF income < 6171.85 THEN Category = B40 ELSE IF 6171.85=<income <15170.35 THEN Category=M40 ELSE Category = T20
IF State = ‘Kedah’ AND IF income < 3180.85 THEN Category = B40 ELSE IF 3180.85=<income <9390.15 THEN Category=M40 ELSE Category = T20
IF State = ‘Terengganu’ AND IF income < 4110.00 THEN Category = B40 ELSE IF 4110.00=<income <9390.15 THEN Category=M40 ELSE Category = T20
IF State = ‘Kelantan’ AND IF income < 6814.58 THEN Category = B40 ELSE M40 ELSE Category = T20
IF State = ‘N.Sembilan’ AND IF income < 4110.00 THEN Category = B40 ELSE IF 4110.00=<income <7622.05 THEN Category=M40 ELSE Category = T20
IF State = ‘Selangor’ AND IF income < 5872.65 THEN Category = B40 ELSE IF 5872.65=<income <12435.15 THEN Category=M40 ELSE Category = T20
IF State = ‘Sarawak’ AND IF income < 3461.75 THEN Category = B40 ELSE IF 3461.75=<income <7622.05 THEN Category=M40 ELSE Category = T20

b) Generating class for overspending: The overspending category was divided into two parts which were 0 and 1. 0 implied that the group fall under not overspending category while 1 was for the group whose total expenditure did exceed the total income. The preparation of overspending class was done by using excel software using the following formula:

IF (Revenue - Spend) <0 THEN category overspending = 0

ELSE category overspending = 1

12 attributes as in Table II were ranked using classifier method to obtain influencing factor. As a result, 8 attribute which produced meaningful reading and ranked higher among the rest was obtained with category (0.19), sex (0.05), education (0.05), ethnic (0.03), marriage (0.02), number of a family (0.02), province (0.018) and state (0.018).

2) Discretization: Discretization is a method which converting continuous data into categorical data [27]. For this method, data form three attributes were processed using depth equal frequency method. Table III shows the description of discretized attribute.

D. Model Development

The classification model was developed using the 10 fold-cross validation using application WEKA (Waikato Environment for Knowledge Analysis). 10 experiments were conducted using each five classification models which were J48, Naïve Bayes, Artificial Neural Network, Swarm Vector Machine and k-Nearest Neighbour. Table IV shows a sample of experimental model that has been conducted using kNN with evaluation parameter as Model number, Number of the fold (training: testing), Accuracy, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), F-measure, Precision, Recall, Kappa statistic, ROC and speed (time taken to build the model). The bold data was best model when compared to all parameters.

The experiment also conducted using other four classification model where the nearest accuracy was form J48 with accuracy of 70%.

TABLE II. ATTRIBUTES DESCRIPTION AND ITS VALUE USED TO DEVELOP THE OVERSPENDING MODEL ON DEMOGRAPHIC

Attributes Name (Variable)	Type	Value
Household Id (HId)	Integer	1, 2, 3 ,…
Number of members in each household (Members)	Integer	
Province (Province)	Integer	1-Peninsular Malaysia 2-Borneo
State (State)	Integer	1-13: Johor, Malacca, Negeri Sembilan, Selangor,….
Ethnic (Ethnic)	1-4	1-Bumiputera, 2-Chiness, 3- India, 4- Other
Education Certificate (Cert)	Integer	1-Tiada Sijil, 2- ,,,,
Marriage Status (Marriage)		1- Marriage, 2- Single Mother 3- Single Father 4- Single
Sex (Sec)	Integer	1- Male 2- Female 3- other
Total Income (total Income)	Real	
Total Expenses (Expenses)	Real	
Amount Over Spending (Oversp)	Real	
Poverty Category (Category)	String	B40, M40, T20

TABLE III. DESCRIPTION OF ATTRIBUTE DISCRETIZATION

Attributes	Type	Range
Number of Member each Household	Integer	1- Members< 2.5 2- 2.5<= Members<4.5 3- Members >=4.5
Total Expenses	Integer	1- Expenses < 1690.01 2- 1690.01 < Expenses<3454. 50 3- 3454.50< Expenses < 4800.59 4- 4800.59< Expenses <5367.67 5- 5367.67<Expenses <6557.03 6- 6557.03<Expenses<7297.58 7- 7297.58 Expenses<8924.1 8- 8924.1< Expenses<9875.87 9- 9875.87<Expenses<10597.08 10-10597.08<Expenses<13169.84

TABLE IV. A SAMPLE EXPERIMENT RESULT OF K-NEAREST NEIGHBOUR CLASSIFICATION MODEL

Model	Fold	Accuracy	MAE	RMSE	F-Measure	Precision	Recall	Kappa statistic	ROC	Speed
1	2	83.31%	0.1679	0.2982	0.91	0.853	0.833	0.1392	0.697	0
2	3	83.96%	0.1611	0.3013	0.802	0.786	0.84	0.2133	0.696	0
3	4	84.12%	0.1487	0.2874	0.819	0.802	0.841	0.3202	0.789	0.01
4	**5**	**84.77%**	**0.1452**	**0.2834**	**0.821**	**0.811**	**0.848**	**0.3089**	**0.781**	**0.01**
5	6	84.45%	0.1457	0.2855	0.822	0.807	0.844	0.3291	0.788	0
6	7	85.18%	0.1478	0.2852	0.825	0.815	0.852	0.328	0.775	0.01
7	8	83.88%	0.1497	0.2912	0.813	0.796	0.839	0.2843	0.779	0.01
8	9	83.96%	0.1483	0.2881	0.812	0.794	0.84	0.2782	0.762	0.01
9	10	84.53%	0.1471	0.2876	0.823	0.81	0.845	0.3253	0.778	0.01

IV. Analysis Result

A. Overspending Phenomenon

Six descriptive Analyses which is distribution income category based on State, Statistical Analysis of Income, Expenditures and Overspending, Income and Overspending based on income category, distribution expenditures of overspending group by type of spending, and statistical analysis of average types of overspending and income category.

1) Distribution of income category based on state: Income data analysis distribution is performed by making SQL directories of household, state, and class ID variables and importing them to Microsoft Excel. Fig. 1 shows the distribution of the income category by state.

The data shows that the population of Sabah, Sarawak and Selangor are relatively high compared to other states. The Federal Territories of Putrajaya, Labuan and Perlis show the least amount of data.

2) Statistical analysis of income, spending and overspending: Table V shows the results of statistical analysis of income, expenditure and overspending among the B40, M40 and T20 classes. The table also shows the percentage distribution of the number of households analysed in this study. 8.5% of the population fall into overspending, where 42% of the households belonged to the B40, 39% to the M40 while the T20 to only 10%. Out of a total of RM91 million monthly income, RM 54 million was spent each month. However, the study found 1027 households suffering from overspending in B40, while 166 in B40 and 35 in T20, which is about 17,2,1 per cent, respectively. The analysis results also show some of the B40 group is spending more than the M40 and T20. The minimum spends per month for the overpaid is 70 cents especially for students. It can be concluded that 83% B40 lifestyle was able to manage their money very well and very little per cent of M40 and T20 fall into overspending. It is undeniable that the B40 group in Malaysia is very wise to save the money they have as only 17% of the B40 group belongs to those who are overspending. Whereas in the M40, only 2.9% overspending and T20 1.3% overspending. However, these percentage overspending population can cause social problems such as theft, bribe or bankruptcy.

The table shows the maximum overspending amount for B40 is RM 5868. It can be concluded that the amount of RM6000 per household is become significant number avoiding household fall into overspending.

3) Income range and overspending by income class: Further analysis on overspending shows in Table VI. The table shows the B40, M40 and T20 income class populations that fall in the overspending category by income range and spending range.

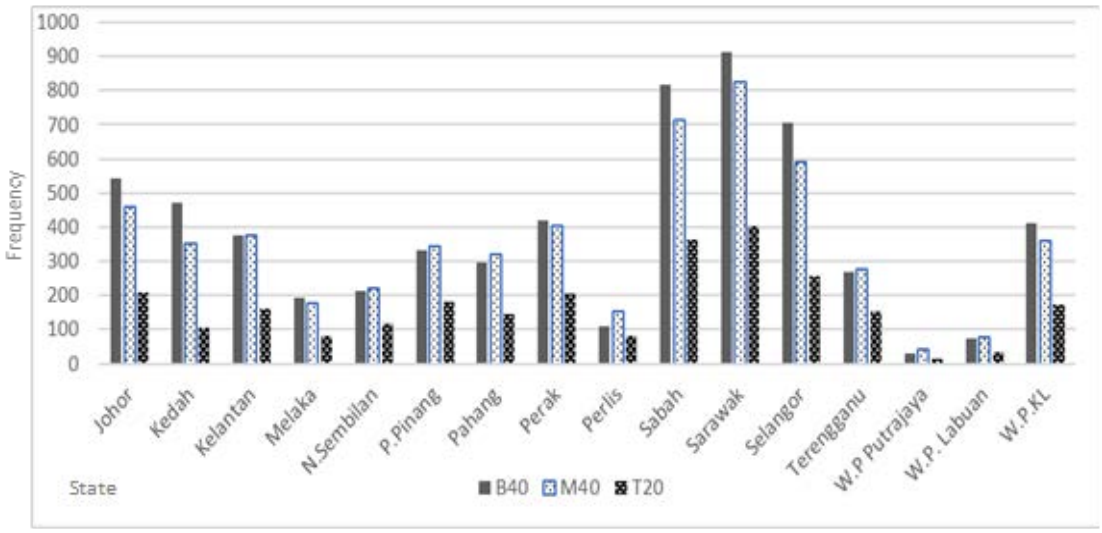

Fig. 1. Frequency of Income Category by State.

TABLE V. Accumulated Amount of Income and Expenditure as Well as Overspending by Income Class

CATEGORY	RESIDENT PERCENTAGE BY CATEGORY		INCOME (RM)	EXPENDITURE (RM)		OVERSPENDING AMOUNT (RM)		EXP>INC
	BIL	(%)		EXPENSE	BALANCE	Min	Max	
B40	6,174	42	18,233,522.61	14,215,885.97	4,017,636.64	0.07	5868.71	**1027**
M40	5,688	39	34,309,736.29	21,353,893.22	12,955,843.07	3.15	10938.27	**166**
T20	2,689	18	38,488,846.37	18,568,726.25	19,920,120.12	92.24	131545.2	**35**
TOTAL	**14,551**	**100**	**91,032,105.27**	**54,138,505.44**	**36,893,599.83**			**1228**

TABLE VI. THE TOTAL EXPENDITURE GROUPS ACCORDING TO THE RANGE OF INCOME AND EXPENDITURE RANGES

Income Range		Expenditure Range		B40	M40	T20
0.00	2,768.10	0.00	5,755.01	769 (75%)		
2,768.10	5,197.70			228 (22%)	72 (40%)	
5,197.70	7,627.30			4	4	
0.00	2,768.10	5,755.01	10,981.86	2		
2,768.10	5,197.70			13	29	
5,197.70	7,627.30			19	23	
7,627.30	10,056.90			3	9	12
0.00	2,768.10	10,981.85	16,208.70			
2,768.10	5,197.70			2		
5,197.70	7,627.30				8	
7,627.30	10,056.90			4		
10,056.90	12,486.50					7
12,486.50	14,196.10					3

The table shows the most populous population of the overspending is income below 2,768. 75% of B40 fall this category, while 22% B40 and 40% M40 fall overspending for income more than RM 2,768.1 and less RM 5,197. The range shows that 97% B40 fall into overspending with income less RM 5197.70.

Data also show 39% of M40 fall into overspending for income less RM 5197.70. Its indicators that basic need of living in Malaysia per household should be less than RM 5500 but nice if RM 6,000. However, these findings are not conclusive because there is a need for more in-depth study of the aspects of spending patterns among Malaysians that need to be studied. The analysis results do not describe the type of expenditure allocated that they belong to the overspending group, which fall into the leisure lifestyle.

The data also shows only 3% of B40 fall overspending more than RM 5,5197.7 as we defined as improper financial management. Seven 7 households fall into this category, most of the area family with 2-5 number of members per household, income between RM3K to RM7K, education is SPM and diploma, 2 females and 5 males, and from Johor, Kedah and Selangor. Similarly, data also shows about 60% of M40 fall into overspending in which income more than RM 5,519.7. Most M40 overspending is from Kedah, Sarawak and Perak, their education level mostly SPM/STPM, they are Bumiputera and Chinese. While for T20, most of them area Bumiputera and India, with education level either SPM, certificate or diploma and they are scattered in Malaysia.

B. Classification Experiment Result

As stated earlier, the aims for mining the data is to discover knowledge on overspending pattern using a classification approach. Table VII shows the summary of classification accuracy result for the five classification models. Where, the bold value showed the best result for each classification model. The accuracy shown here is represented the best model obtained at which fold of training vs. testing data. SVM has proven to be the best model for household overspending model amongst B40, M40 and T20 with accuracy of 89.17% followed by J48, ANN, kNN and Bayes with accuracy of 88.84%, 86.97%, 84.77% respectively. Even though SVM shows the best accuracy, J48 classification model was selected for rule generation. This is due to the ability of J48 to presenting the model in the form of rule which make it easier for knowledge discovery.

Tables VIII to X show the overspending rules generated from six types of attributes for B40, M40 and T20 group respectively. The rules were extracted from decision tree developed by J48 algorithm which represent the important in sequence.

From the rules generated, it can be seen that for B40 group that live in Melaka, Perlis, Perak and Sarawak, over spending happened. For B40 that live in Perlis, Federal Territories, Labuan, Penang, Johor and Selangor, overspending happened if expenditure below RM 3952.62. Overspending also happen among B40 if total expenditure is between RM 3952.62 and RM 6557.03 for resident of Johor having fewer or 4 children. In Terengganu, the B40 group said to be overspend with total expenditure of less than RM4800.59. In Malacca, B40 group considered as overspent when their spending reached RM5367.57. Lastly, overspending would happen if B40 lives in rural area of Kedah.

In Perak, Sarawak, Perlis, Labuan and Selangor, M40 who spend more than RM 6557.03 considered as overspent. In Melaka, the M40 overspend when the total expenditure reached RM 5367.67. In Labuan, the M40 overspend when the total expenditure reached RM 10350.4. However, for M40 who lives in Kedah with total expenditure between RM7297 and RM10350.4 they considered to be in overspend category. In Kuala Lumpur and Terengganu, the M40 group overspend when their total expenditure exceeds RM4800, especially for those living in the city. These rues can be seen in Table IX.

Table X shows the overspending rules for T20 group. T20 group considered to be overspent when their total spend exceeds RM 6557.03. The overspending T20 is among those who lives in Negeri Sembilan and Pahang. In Malacca, T20 considered to be overspent when total expenditure exceeded RM9875.87. Lastly, in Labuan, the T20 group is overspent when the total expenditure exceeds RM10350.4.

TABLE VII. COMPARISON ACCURACY RESULT AMONGST FIVE CLASSIFICATION MODEL

Model	Folds	Accuracy %				
		J48	ANN	Bayes	SVM	kNN
1	2	87.30	85.59	83.06	87.95	82.74
2	3	87.87	87.13	83.55	89.09	83.63
3	4	88.03	**86.97**	83.14	88.52	83.06
4	5	88.60	85.83	83.39	88.84	**84.77**
5	6	88.68	86.81	82.74	89.09	84.12
6	7	88.60	86.89	83.39	88.52	84.61
7	8	88.68	86.97	82.98	**89.17**	83.55
8	9	88.60	85.91	83.63	89.09	83.47
9	10	**88.84**	86.56	**83.63**	89.09	83.88
Average Accuracy		88.36	86.52	83.28	88.82	83.76

TABLE VIII. OVERSPENDING FOR B40

No	Rules
1	If Expenses <= RM6557.03 AND State is Kedah AND Strata is Rural
2	If Expenses <= RM4800.59 AND State is Terengganu
3	If State is Kedah AND Expense between RM7297.59 AND RM10597.08 or more than RM13169.84
4	If Expenses <=RM6557.03 AND State is N. Sebilan AND household members <=4
5	If Expenses <=RM6557.03 AND State is Perlis OR W.P.KL Or W.P.Labuan Or Pulau Pinang OR Johor OR Selangor
6	If Expenses <=RM6557.03 AND State is Sabah and Ethnic is India OR Chinese OR Bumiputra
7	If Expenses >RM6557.03 AND State is W.P.KL AND household memebers <=4
8	If Expenses >RM6557.03 AND State is Johor
9	If Kedah AND Expense >13169.84
10	IF Expenses <=RM5367.67 AND State is Melaka

TABLE IX. OVERSPENDING FOR M40

No	Rules
1	IF Expenses <= RM6557.03AND State is Kelantan OR Perak OR Sarawak OR Pahang
2	IF Expenses <= RM6557.03 AND State is Kedah AND Strata is Urban
3	IF Expenses <= RM6557.03 AND State is N. Sembilan AND Member of Household is >4
4	IF Expenses >RM4800.59 AND State is Terengganu
5	IF Expenses <= RM6557.03 AND State is Sabah AND Ethnic is Others
6	IF Expenses between RM5367.67 and RM6557.03 AND State is Melaka
7	IF Expenses between RM6557.03 and RM9875.87 and State is Sabah
8	IF Expenses > RM6557.03 and State is W.P.KL AND household members >4
9	IF Expenses between RM6557.03 and RM10350.48 AND state is Pulau Pinang
10	IF Expenses between RM6557.03 and RM8924.1 AND State is Terengganu
11	IF Expenses > RM6557.03 AND State is Perak OR Sarawak OR Perlis OR W.P.Labuan OR Selangor
12	IF Expenses between RM7297.58 and 10597.08 AND State is Kedah
13	IF Expenses >RM10597.08 AND State is Kedah

TABLE X. OVERSPENDING FOR T20

No	Rules
1	IF Expenses > RM6557.03 AND state is Kelantan OR Pahang OR N.Sembilan
2	IF Expenses > RM9875.87 AND State is Sabah
3	IF Expenses is >RM10350.48 AND State is Pulau Pinang
4	IF Expenses >RM 8924.1 AND State is Terengganu

V. CONCLUSION

Studies have shown descriptive analysis results and analytical data analysis using the J48 classification method to produce demographic-based and overspending rules based on expenditure type. Descriptive analysis showed the distribution statistics of the B40, M40 and T20 groups by state. Comparative analysis of the two variables can be performed individually to compare or produce specific patterns. Similarly, a descriptive analysis of overspending on expenditure type shows the average distribution of expenditure types for B40, M40 and T20.

There are six attributes that influence most to overspend which are state, race, income, strata, number of households and categories. The rules can determine the exact rule for overspending for each state. The rules show the attractive features of the overspending when the total expenditure is less than or above RM 6557 per state followed by other features such as race, strata and household numbers. For example, in Kuala Lumpur, the M40 group with more than four children is facing overspending with a cost of over RM 6557. With this model, it can help the B40 category not only depend on the amount of income but also need to look at the number of household aspects. The model also proved that the number of household member can be one of the variable in identifying poverty category as B40, M40 or T20.it can be seen that the higher the number of household member, the higher the total expenses. Moreover, ethnic also become one of the overspending factor which only can be applied in Sabah.

This paper has shown how data analytic help in identifying knowledge of attributes that influence the category of poverty based on demography. However, the rules produced were based on the B40, M40 and T20 data which only covering 20 fractions of the actual number of questionnaires conducted by DOSM. The findings may be inaccurate due to the limited amount of data that reflects the actual occupation and population of Malaysians. This study able to show how data science or analytical data in the domain of data mining able to find more detailed knowledge than existing data sources. The results of the descriptive data section and the analysis of the overspending models in B40, M40 and T20 provided clear picture of analytical data capabilities in the pursuit of more detailed knowledge to assist authorities in making decisions or planning strategic plans for income and expenditure management in Malaysia. At the end of this study, factors of income, state, and number of children/members per house hold were found to be among most influence factors in determining Malaysian Family wellbeing index. However, this paper which focusing on demographic data could not assist in identifying lifestyle. Further research focusing on development overspending model based on type of expenses such as total food and transportation are more beneficial to identify the lifestyle. Therefore, study on type of expenses that have highest amount of expenses can be done in the future. This thus can urge these group to minding their expenses in that base so that they would not fall into overspending category.

ACKNOWLEDGMENT

This study is supported by the Universiti Kebangsaan Malaysia (UKM) and funded by research grant (DCP-2017-015/4).

REFERENCES

[1] Department of Statistic Malaysia, “Malaysian Well-Beng Index 2018”. [Press Release]. December 2019. Available at https://www.dosm.gov.my/v1/index.php?r=column/pdfPrev&id=UHpYdlBUZFhUU0RoTTJidFc0SWwrZz09 (Accesseed 10 July 2020).

[2] Gui, M., & Büchi, M. (2019). From Use to Overuse: Digital Inequality in the Age of Communication Abundance. Social Science Computer Review, 089443931985116. doi:10.1177/0894439319851163.

[3] Carick, R. "Why household overspending is worse than the federal deficit," The Globe and Mail, March 17, 2019 . [Online]. Available: https://www.theglobeandmail.com/investing/personal-finance/article-why-household-overspending-is-worse-than-the-federal-deficit/, [Accessed June 14, 2020].

[4] Zakaria, R.H., Jaafar, N.I.M. and Ishak, N.A.., "Household debt decision: Poverty or psychology?" International Journal of Business and Society, vol. 18, pp. 515-532.

[5] Duffin, E., "Number of personal bankruptcy filings nationwide in the U.S. 2000-2019," Statista. [Online] Available: https://www.statista.com/statistics/817911/number-of-non-business-bankruptcies-in-the-united-states/, [Accessed June 14, 2020].

[6] Bankruptcy Statistic Disember 2018. Official Portal Malaysian Department of Insolvency [Online]. Available: http://www.mdi.gov.my/index.php/ms/about-us/resources/statistics/bankruptcy/1390-bankruptcy-statistic-disember-2018. [Accessed April 15, 2020].

[7] Mozaher, K., Yokota, F., Nishikitani, M., Islam, R., Kitaoka, H., Fukuda, A., Ahmed, A., Muske, G. and Winter, M. "Factors associated to online shopping at the bopcommunity in rural Bangladesh". (IJACSA) International Journal of Advanced Computer Science and Applications,vol. 8, no. 10, 2017.

[8] Rashid, N.K.A., Sulaiman, N.F.C. and Rahizal, N.A., "Survivability through basic needs consumption among muslim households B40, M40 and T20 income groups," Pertanika Journal of Social Sciences & Humanities, vol. 26, no. 2, 2017.

[9] Department of Statistics Malaysia. Household Expenditure Survey Report 2019 [Press releaese] July 2020. Available at https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=323&bul_id=c3JpRzRqeTNPamMxL1FpTkNBNUVBQT09&menu_id=amVoWU54UTl0a21NWmdhMjFMMWcyZz09. (Accessed 13 July 2020).

[10] Sevakula, R.K., Au ‐ Yeung, W.T., Jagmeet, P.S., Heist, E.K., Isselbacher, E.M, and Armoundas, A.A, "State ‐ of ‐ the ‐ art machine learning techniques aiming to improve patient outcomes pertaining to the cardiovascular system," Journal of the American Heart Association, vol. 9, no.4, 2020.

[11] Abuhamad, H.I.S., Bakar, A.A., Zainudin, S., Sahani, M. and Ali, Z.M., "Feature selection algorithms for Malaysian dengue outbreak detection model," Sains Malaysiana, vol. 46, no. 2, pp. 255-265, 2017.

[12] Nazri, M.Z.A., Ghani, R.A., Abdullah, S., Ayu, M., and Samsiah, R.N., "Predicting academician publication performance using decision tree," International Journal of Recent Technology and Engineering, vol. 8, no.2 (Special Issue), pp.180-185, 2019.

[13] Gopagoni, D.R. and Lakshmi P.V., "Automated machine learning tool: the first stop for Data Science and Statistical Model Building," IJACSA, vol. 11, no. 20, pp. 410-418, 2020.

[14] Azlan, W.A.W., Liew, S.H., Choo, Y.H., Zakaria, H. and Low, Y.F., "Wavelet feature extraction and J48 decision tree classification of auditory late response (ALR) elicited by transcranial magnetic stimulation," ARPN J. Eng. Appl. Sci, vol. 11, no. 10, pp.6319-6323, 2016.

[15] Kaunang, F. J., "Application of j48 decision tree algorithm for analyzing poverty level in Indonesia," Cogito Smart Journal, vol. 4, no. 2, pp. 348-357, 2018.

[16] Niu, T., Chen, Y., and Yuan, Y., "Measuring urban poverty using multi-source data and a random forest algorithm: A case study in Guangzhou," Sustainable Cities and Society, vol 54, 2020.

[17] Jarrah, M and Salim, N., "A recurrent neural network and a discrete wavelet transform to predict the Saudi stock price trends," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 10, no. 4, 2019.

[18] Kahar, S.A, Sheikh, N., Adnan, N., Iqbal, S., Rehman, A., Kahar, A. U., Kakaar, B.A, Kakar, H.A. and Khan, B., "Artificial neural network based weather prediction using Back Propagation Technique," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 9, no. 8, pp. 462-270, 2018.

[19] Khan, Y., Shafiq, S, Naeem, A., Hussain, S., Ahmed, S., and Safwan, N., "Customers churn prediction using artificial neural networks (ANN) in telecom industry," International Journal of Advanced Computer Science and Applications, vol. 10, no. 9, 2019.

[20] Babenko, B., Hersh, J., Newhouse, D., Ramakrishnan, A., and Swartz, T., "Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in Mexico," arXiv:1711.06323v1 [stat.ML], 2017.

[21] Hemmatian, F. and Sohrabi, M.K. "A survey on classification techniques for opinion mining and sentiment analysis," Artif Intell Rev, vol. 52, pp. 1495–1545, 2017.

[22] Shih, A., Choi, A., and Darwiche, A., "Compiling bayesian network classifiers into decision graphs," Proceedings of the AAAI Conference on Artificial Intelligence, no.33, pp. 7966–7974, 2019.

[23] Letham, B., Karrer, B., Ottoni, G., and Bakshy, E., "Constrained bayesian optimization with noisy experiments," Bayesian Analysis, vol. 14, no. 2, pp. 495–519, 2019.

[24] Redjeki, S., Guntara, M. and Anggoro, P., "Naive bayes classifier algorithm approach for mapping poor families potential," (IJARAI) International Journal of Advanced Research in Artificial Intelligence, vol. 4, no.12, pp.29-33, 2015.

[25] Aggarwal, C.C and Yu, P.S. "A condensation approach to privacy preserving data mining," Externl Database Technology (EDBT), pp 183-199, 2004.

[26] Laporan penyiasatan perbelanjaan isi rumah," Jab. Prnk. MALAYSIA, 9 October 2017 [Press reelased] Available at https://www.dosm.gov.my/v1/index.php?r=column/pdfPrev&id=V3R4SHBmeUhDUWlETjZUbXdWbXB2dz09. (accesssed 15 July 2020).

[27] J Stańczyk, U., Zielosko, B., and Baron, G., "Discretisation of conditions in decision rules induced for continuous data," PLoS ONE, Vol 15 No 4, 2020.