IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 May 2022, accepted 27 June 2022, date of publication 30 June 2022, date of current version 8 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3187564

APPLIED RESEARCH

Machine Learning Models for Predicting Financially Vigilant Low-Income Households

RATHIMALA KANNAN[1], (Senior Member, IEEE), KHOR WOON SHING[2], KANNAN RAMAKRISHNAN[3], (Senior Member, IEEE), HWAY BOON ONG[4], AND ANDRY ALAMSYAH[5], (Member, IEEE)

[1]Department of Information Technology, Faculty of Management, Multimedia University, Cyberjaya, Selangor 63100, Malaysia
[2]Faculty of Management, Multimedia University, Cyberjaya, Selangor 63100, Malaysia
[3]Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor 63100, Malaysia
[4]Department of Economics, Faculty of Management, Multimedia University, Cyberjaya, Selangor 63100, Malaysia
[5]School of Economics and Business, Telkom University, Bandung, West Java 40257, Indonesia

Corresponding author: Rathimala Kannan (rathimala.kannan@ieee.org)

ABSTRACT The COVID-19 pandemic has adversely affected households' lives in terms of social and economic factors across the world. The Malaysian government has devised a number of stimulus packages to combat the pandemic's effects. Stimulus packages would be insufficient to alleviate household financial burdens if they did not target those most affected by lockdowns. As a result, assessing household financial vigilance in the case of crisis like the COVID-19 pandemic is crucial. This study aimed to develop machine learning models for predicting and profiling financially vigilant households. The Special Survey on the Economic Effects of Covid-19 and Individual Round 1 provided secondary data for this study. As a research methodology, a cross-industry standard process for data mining is followed. Five machine learning algorithms were used to build predictive models. Among all, Gradient Boosted Tree was identified as the best predictive model based on F-score measure. The findings showed machine learning approach can provide a robust model to predict households' financial vigilances, and this information might be used to build appropriate and effective economic stimulus packages in the future. Researchers, academics and policymakers in the field of household finance can use these recommendations to help them leverage machine learning.

INDEX TERMS Economic stimulus packages, financial vigilance, low-income households, machine learning, prediction.

I. INTRODUCTION

The COVID-19 pandemic had triggered a global health crisis and caused many countries to implement lockdowns that disrupted economic activities worldwide [1]. Economic lockdowns will cause an income shock, especially in low income households [2]. Low income households do not have sufficient excess income to save. Therefore, when they lose their source of income, they will not be able to pay for their daily essentials such as food and shelter.

The Malaysian government allocated several economic stimulus packages to ease the financial burden of low income households [3]. However, the problem with the government's stimulus packages is its lack of focus to accurately target the households impacted by the COVID-19 pandemic [4]. The COVID-19 pandemic had affected low-income households as well as those in the M40 and T20. Take for instance, the aviation industry. Many pilots, cabin and ground crew, were laid off due to travel restrictions and closed borders. These households were previously not in the B40 category, but they had lost their source of income. Many households in the tourism related industry also faced a similar fate. Without any source of income, it is difficult to be financially vigilant during these trying times of the pandemic.

The Department of Statistics Malaysia (DOSM) conducted two special surveys to understand how the pandemic affected Malaysian households. The first survey was conducted between 23rd and 31st March 2020 prior to

The associate editor coordinating the review of this manuscript and approving it for publication was Victor Sanchez.

implementing the movement control order (MCO) and received 168,145 responses. The survey consists of the respondents' feedback based on their opinions on economics, employment, and spending patterns. The second survey was conducted between 10th to 24th April 2020, after the first MCO, and received 41,386 responses. These two survey data were made open to the researchers [5] and hence this study is motivated to extract the insights further.

This paper aimed to apply data analytics and machine learning techniques on the first survey data to extract insights that would be helpful to gain a better understanding of COVID-19 impacts on the financial vigilance of households in Malaysia. More precisely, this research attempts develop machine learning models for predicting financially vigilant households in the face of the covid-19 pandemic. The outcome from this study would be useful in determining the impact of pandemic on households, as well as providing machine learning models to predict households' financial vigilance and develop a future economic stimulus package.

II. RESEARCH BACKGROUND

A. ECONOMIC STIMULUS

An economic stimulus is an economic policy tool put in place by the government to cushion adverse economic effects and help economic recovery. There are two types of economic stimulus, namely, monetary stimulus and fiscal stimulus. Monetary stimulus refers to the nation's central bank's policy tool directed at the domestic interest rate to influence consumption and investment. For example, an expansionary monetary policy is implemented during an economic downturn to reduce the cost of funds and encourage consumption.

Fiscal stimulus requires an increase in government spending to bring the economy out of the crisis by providing incentives and tax rebates. Fiscal stimulus refers to the actions of governments with lowering the tax rates, increase subsidies and allowances to boost household spending [6]. Most countries have implemented monetary and fiscal stimulus to apprehend economic slowdown due to lockdowns during the COVID-19 pandemic.

B. STIMULUS PACKAGES IN MALAYSIA

Without a source of income, many households would quickly lose access to essential needs like food or housing [7]. In order to lessen their financial burden, the Malaysian government has created a series of stimulus packages. The stimulus package aims to help households and businesses to sustain economic lockdowns [8]. The first stimulus package was aimed to support the SMEs, tourism and hospitality sectors. It also focused on promoting high value-added investments in both the public and private sectors. The first stimulus package consists of loan reschedule for businesses, RM2 billion Special Relief Facility for SMEs, service tax exemption for hotels and tourism industry, special allowance to health workers, and electricity bill discount of 15%.

The second stimulus package is named the Prihatin Rakyat Economic Stimulus Package and implemented on 27th March 2020. This stimulus package was created to support the M40 and B40 income groups in easing the financial burden of households and individuals in a direct cash transfer programme. The third stimulus package was released to support SMEs and was called the Prihatin SME Economic Stimulus Package. This package was aimed to secure employees from getting terminated, which otherwise will lead to job loss and a high unemployment rate [5].

The COVID-19 Pandemic has highlighted the vital role of digital technology, innovations, and services in allowing government, companies, and society to work during crises. In addition to ensuring stability and connectivity, digitalisation lays the groundwork for a more sustainable and inclusive economic transition. COVID-19 recovery stimulus plans have been tracked by the World Bank's Digital Development Global Practice worldwide, with a specific emphasis on possible avenues for integrating digital ICT technology and digital services as part of COVID recovery efforts [9].

C. DATA ANALYTICS AND MACHINE LEARNING

Most governments in industrialised nations have identified using and enhancing information and communication technology to enhance public sector services (e-Government) as a critical goal [8], [10]. Data analytics is an important part of extracting hidden knowledge from data gathered from various sources. It refers to the scientific methods, processes and machine learning techniques used to extract huge databases. The approaches analyse data in real-time to find non-linear correlations and causal effects commonly hidden throughout datasets [11]. Machine learning is ideally aligned with the breadth of the next generation of government, Government 3.0, which examines all-new options to address any difficulty confronting modern societies by using new technology for data-driven decision making [12], [13].

Despite a large number of studies on Machine Learning applications in the literature, the government domain has received relatively little attention [8]. For example, a study developed a general profile for users of e-government services based on demographic, cognitive, and psychographic characteristics [14]. The authors had applied popular classification techniques such as neural networks, decision trees and rule induction. Another research employed machine learning approaches to assist government agencies in dealing with the arduous task of receiving and categorising customer complaints [15]. Finally, based on Artificial Intelligence (AI) and Machine Learning (ML) algorithms, a recent study used available government data to forecast the vulnerability of specific enterprises to economic catastrophe [16].

Reference [10] applied several data mining algorithms to create machine learning models that can accurately forecast the e-Government rankings of 192 United Nations and identify the variables that influence those rankings. Reference [17] developed a COVID-19 Vulnerability Index (C19VI) by applying machine learning based on sociodemographic COVID-19-specific themes. Based on C19VI and census data, the authors found that the epidemic has disproportionately affected racial minorities and the underprivileged. Based on previous studies, we can see that machine learning techniques can be utilized on government data to extract hidden information that can be helpful in data-driven decision-making and developing new policies for the citizens' well-being [17].

III. RESEARCH METHODOLOGY

This research applied the Cross-Industry Standard Process-Data Mining (CRISP-DM) to plan and execute the project. CRISP-DM is the de-facto standard for applying data mining projects and is an industry-independent process model [18]. This methodology consists of 6 phases; business understanding, data understanding, data preparation, modelling, evaluation and deployment. Fig. 1 illustrates the research methodology of this study.

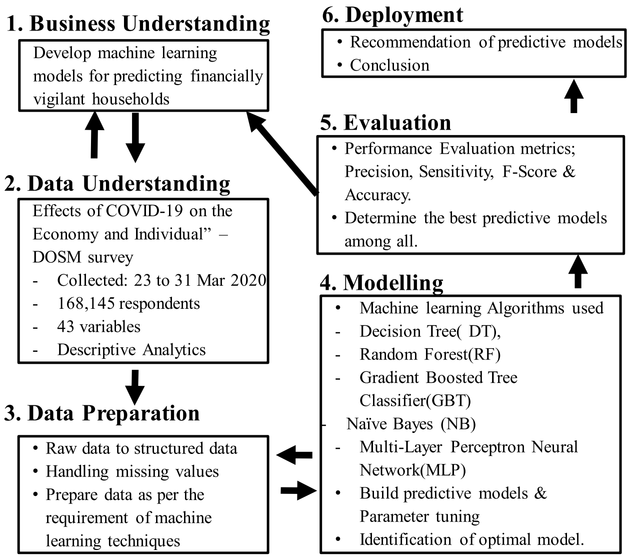
1. Business Understanding
Develop machine learning models for predicting financially vigilant households
2. Data Understanding
Effects of COVID-19 on the Economy and Individual" – DOSM survey
- Collected: 23 to 31 Mar 2020
- 168,145 respondents
- 43 variables
- Descriptive Analytics
3. Data Preparation
• Raw data to structured data
• Handling missing values
• Prepare data as per the requirement of machine learning techniques
4. Modelling
• Machine learning Algorithms used
- Decision Tree(DT),
- Random Forest(RF)
- Gradient Boosted Tree Classifier(GBT)
- Naïve Bayes (NB)
- Multi-Layer Perceptron Neural Network(MLP)
• Build predictive models & Parameter tuning
• Identification of optimal model.
5. Evaluation
• Performance Evaluation metrics; Precision, Sensitivity, F-Score & Accuracy.
• Determine the best predictive models among all.
6. Deployment
• Recommendation of predictive models
• Conclusion

FIGURE 1. **Research methodology.**

A. BUSINESS UNDERSTANDING

The business understanding phase focuses on the project's objectives and requirements that they want to achieve in the project. This phase will also determine the purpose of the machine learning perspective to achieve the project's goals [18]. In this phase, the study's objective was to extract insight from the survey data and determine the effect of COVID-19 on the economy and individuals in Malaysia by applying machine learning techniques. This research also attempts to provide a general profile of individuals who are or are not financially prepared to deal with pandemic-like situations.

B. DATA UNDERSTANDING

This step requires gathering data from many sources, examining and characterising it, and ensuring data quality. The secondary dataset utilised in this study comes from DOSM, a special survey on the effects of Covid-19 on the economy and individuals that was collected from March 23rd to March 31st, 2020. The original survey had three modules and twenty-one questions, with a total of 168,182 participants aged 15 and up. However, the dataset obtained from DOSM consists of 168,145 records/rows and 43 variables/columns.

The original dataset was in the Malay language. For the study purpose, the whole dataset had been translated into the English language. After the translation, the rename nodes redefine and shorten the variable names to be more understandable. There are 6,365 rows of missing records found in the dataset. These 6,365 missing records had been filtered and separated into the new table for visualisation purposes. This study used the open-source software KNIME Analytics Platform to execute the machine learning project.

C. DATA PREPARATION

In this third phase, defining inclusion and exclusion criteria should be used to choose data. The dataset is noisy with missing values and related issues, so considerable time is spent cleaning the dataset before applying descriptive analytics techniques. This step requires reformatting the dataset, such as string to number to prepare for algorithms' use [18].

A new attribute, income category, was derived from the "Monthly Income" attribute, which groups the income into B40, M40 and T20. B40 income category will be the non-working respondents, <RM2,500, and RM2500-RM4000 for their monthly income. M40 income category will be the respondents who earn RM4,001-RM6,000, RM6,001-RM8,000, and RM8,001-RM9,000 for their monthly income. Those respondents who earn more than RM9,000 will be categorised as "T20". Those who did not provide their monthly income information will be classified as "NA" (Not Applicable/no income information). The income group will be slightly different from reality due to the limitation of the dataset. "NA " income category had been filtered (56,459 respondents) from the modelling since the monthly income information is important to the modelling purpose. Final dataset consists of 105,321 records.

D. MODELING

In this phase, different machine learning algorithms will be applied to achieve the objectives of the study. Five machine learning algorithms: Decision tree, Random Forest, Naïve Bayes, Gradient boosted tree, and Multi-layer perceptron neural network (MLP) were applied to classify citizens' who are/are not financially prepared to cope the Pandemic [8]. When applying each machine learning algorithm, the optimal model was determined by varying the parameters such as partitioning ratio of training and testing data, quality measure: Gini index, information gain and gain ratio.

Gini index also known as Gini coefficient calculate the degree of likeliness of a variable that is wrongly classified during the random selection and a difference of Gini coefficient. It provides binary splitting, either "success" or "fail" outcomes when it comes to categorical values [19].

$$Gini\,(P) = \sum_{i=1}^{n} p_i\,(1 - p_i) = 1 - \sum_{i=1}^{n} (p_i)^2$$

where

$$P = (p_1, p_2, \ldots p_n)$$

$\mathrm{p_i}$ = probability of classification on object to a class

Gain ratio also known as uncertainty coefficient standardises attributes' information gain against degree of uncertainty (entropy). When the amount of entropy is large, the gain ratio will be low and vice versa [19].

$$\begin{aligned} GainRatio &= \frac{InformationGain}{Entropy} \\ InformationGain &= Entropybeforesplitting \\ &\quad - Entropyaftersplitting \end{aligned}$$

E. EVALUATION

In this phase, various machine learning models developed in the modelling phase will be evaluated based on the standard performance metrics accuracy, and f-measures. Finally, the optimal model of each algorithm will be chosen and the results will be combined to form a comparison table. This is to make sure each algorithms' performance is evaluated and proceed to the deployment phase.

F. DEPLOYMENT

This is the last phase of CRISP-DM. Technically the best machine learning model determined from the evaluation phase will be deployed with the real dataset. This step will also review whether the models achieved the project's goals and requirements [18].

IV. RESULTS AND DISCUSSION

In this section, results obtained from descriptive analytics, model optimisation and findings will be discussed. First, it helps to understand the characteristics of each respondent and the relationship between the variables. As mentioned previously, 6,365 respondents did not complete their survey from variable 12, "Working experience/Total Working Years." Therefore, the descriptive analytics for the complete survey (161,780 respondents) is given below.

A. DESCRIPTIVE ANALYTICS

Table 1 presents the demographics of the respondents from the descriptive analytics for the entire survey, which contains 161,780 records with 41 variables. The majority of the respondents came from Selangor (27%), followed by Johor (21%), Kuala Lumpur (9%) and other states. There are 95,158 females (59%) and 66,622 males (41%) respondents' opinions included in this analysis. Besides, most of the respondents were from the 35 to 44 years old age group (37%) followed by the age group from 25 to 34 years old (35%), from 45 to 54 years old (16%) and minor from other age groups (12%).

Most of the respondents stated that their financial vigilance (savings) was sustainable for their daily expenses for about 2 to 4 weeks (47,273) and 40,947 respondents replied that their savings were sufficient to support their daily expenses for 4 to 8 weeks. 37,015 respondents stated that their savings were sufficient in less than two weeks. 77% of the respondents revealed that their financial vigilance period is only for less and equal to 8 weeks. 55% (88,895) responded that they did not prepare financially to cope with the pandemic. It also showed that the majority of the respondents were B40 income group (49%), followed by no information (35%), M40 (13%) and T20 (3%).

TABLE 1. Demography description of the data set.

Variable	Frequency	Percentage
State		
Selangor	43,560	26.93%
Johor	33,341	20.61%
Kuala Lumpur	13,787	8.52%
Sabah	11,359	7.02%
Perak	7,821	4.83%
Negeri Sembilan	6,969	4.31%
Kedah	6,851	4.23%
Pahang	6,110	3.78%
Sarawak	5,962	3.69%
Kelantan	5,161	3.19%
Pulau Pinang	5,031	3.11%
Putrajaya	4,813	2.98%
Melaka	4,728	2.92%
Terengganu	4,378	2.71%
Perlis	1,116	0.69%
Labuan	793	0.49%
Gender		
Female	95,158	58.82%
Male	66,622	41.18%
Age Group		
35-44	60,383	37.32%
25-34	56,482	34.91%
45-54	26,017	16.08%
15-24	10,537	6.51%
55-64	7,677	4.75%
>65	684	0.42%
Ethnic group		
Malay	131,095	81.03%
Native of Sabah/Sarawak	13,360	8.26%
Chinese	11,850	7.32%
Indian	4,331	2.68%
Others Native	908	0.56%
Non-Citizen	200	0.12%
Other citizens	36	0.02%
No. Household Members		
1-5	109,365	67.60%
6-10	49,528	30.61%
11-15	2,407	1.49%
16-20	290	0.18%
>21	190	0.12%

TABLE 2. Identification of optimal model for decision tree.

Partitioning	Quality Measure	Target	F-measure	Accuracy
50:50	Gini Index	Yes	0.6168	0.6707
		No	0.7112	
	Gain Ratio	Yes	0.6262	0.6820
		No	0.7232	
55:45	Gini Index	Yes	0.6223	0.6773
		No	0.7183	
	Gain Ratio	Yes	0.6167	0.6834
		No	0.7303	
60:40	Gini Index	Yes	0.6271	0.6792
		No	0.7186	
	Gain Ratio	Yes	0.6236	0.6804
		No	0.7223	
65:35	Gini Index	Yes	0.6227	0.6788
		No	0.7203	
	Gain Ratio	Yes	0.6240	0.6831
		No	0.7261	
70:30	Gini Index	Yes	0.6306	0.6795
		No	0.7169	
	Gain Ratio	Yes	0.6204	0.6799
		No	0.7233	
75:25	Gini Index	Yes	0.6265	0.6802
		No	0.7204	
	Gain Ratio	Yes	0.6294	0.6855
		No	0.7268	
80:20	Gini Index	Yes	0.6228	0.6777
		No	0.7186	
	Gain Ratio	Yes	0.6292	0.6836
		No	0.7241	
85:15	Gini Index	Yes	0.6290	0.6834
		No	0.7238	
	Gain Ratio	Yes	0.6317	0.6887
		No	0.7305	
90:10	Gini Index	Yes	0.6305	0.6819
		No	0.7208	
	Gain Ratio	**Yes**	**0.6324**	**0.6896**
		No	**0.7313**	

TABLE 3. Identification of optimal model for random forest.

Partitioning	Split Criterion	Target	F-measure	Accuracy
50:50	Gini Index	Yes	0.7140	0.7450
		No	0.7700	
	Information Gain Ratio	Yes	0.7070	0.7435
		No	0.7719	
55:45	Gini Index	Yes	0.7164	0.7472
		No	0.7719	
	Information Gain Ratio	Yes	0.7072	0.7427
		No	0.7705	
60:40	Gini Index	Yes	0.7180	0.7480
		No	0.7722	
	Information Gain Ratio	Yes	0.7098	0.7450
		No	0.7726	
65:35	Gini Index	Yes	0.7133	0.7464
		No	0.7726	
	Information Gain Ratio	Yes	0.7071	0.7435
		No	0.7718	
70:30	Gini Index	Yes	0.7196	0.7503
		No	0.7749	
	Information Gain Ratio	Yes	0.7133	0.7494
		No	0.7774	
75:25	Gini Index	Yes	0.7166	0.7482
		No	0.7734	
	Information Gain Ratio	Yes	0.7102	0.7470
		No	0.7755	
80:20	Gini Index	Yes	0.7178	0.7516
		No	0.7782	
	Information Gain Ratio	Yes	0.7151	0.7512
		No	0.7792	
85:15	Gini Index	Yes	0.7230	0.7530
		No	0.7771	
	Information Gain Ratio	Yes	0.7111	0.7483
		No	0.7770	
90:10	Gini Index	Yes	0.7192	0.7495
		No	0.7740	
	Information Gain Ratio	**Yes**	**0.7191**	**0.7545**
		No	**0.7820**	

Moreover, there were 41,918 respondents (26%) who had to work from home, 26,899 respondents (17%) not working, 11,561 respondents (7%) responded that they were still working but their working hours had decreased and 7,998 respondents (4.9%) had lost their job due to COVID-19 pandemic. Furthermore, 47,173 respondents' (29%) income was not affected by the pandemic and 30,903 (19%) had an income reduction. Besides that, 111,949 respondents (69%) were ready for Movement Control Order (MCO) and 49,831 respondents (31%) were not ready for MCO.

It also found that COVID-19 had an impact on households' lifestyles. Before the COVID-19 pandemic, 44,936 respondents often travelled due to their business or job requirements, but the COVID-19 pandemic had stopped 71,782 respondents from travelling. Some 60,399 respondents travelled for personal reasons but 65,022 stopped travelling for personal reasons due to the pandemic. Next, most of the respondents

TABLE 4. Identification of optimal model for Naïve Bayes.

Partitioning	Target	F-measure	Accuracy
50:50	Yes	0.6761	0.7156
	No	0.7465	
55:45	Yes	0.6777	0.7157
	No	0.7457	
60:40	Yes	0.6721	0.7112
	No	0.7420	
65:35	Yes	0.6802	0.7162
	No	0.7449	
70:30	Yes	0.6825	0.7198
	No	0.7492	
75:25	Yes	0.6786	0.7157
	No	0.7452	
80:20	Yes	0.6715	0.7114
	No	0.7427	
85:15	Yes	0.6725	0.7075
	No	0.7358	
90:10	**Yes**	**0.6841**	**0.7215**
	No	**0.7510**	

TABLE 5. Identification of optimal model for gradient boosted tree.

Partitioning	Target	F-measure	Accuracy
50:50	Yes	0.6947	0.7515
	No	0.7904	
55:45	Yes	0.6906	0.7514
	No	0.7922	
60:40	Yes	0.6954	0.7545
	No	0.7943	
65:35	Yes	0.6933	0.7538
	No	0.7944	
70:30	Yes	0.6933	0.7525
	No	0.7926	
75:25	Yes	0.6972	0.7558
	No	0.7954	
80:20	**Yes**	**0.6973**	**0.7589**
	No	**0.7997**	
85:15	Yes	0.6997	0.7593
	No	0.7992	
90:10	Yes	0.6947	0.7521
	No	0.7914	

TABLE 6. Identification of optimal model for neural network (MLP).

Normalization	Partitioning	Target	F-measure	Accuracy
Min-Max	50:50	Yes	0.6495	0.7147
		No	0.7595	
Z-Score		Yes	0.6643	0.7384
		No	0.7857	
Min-Max	55:45	Yes	0.6604	0.7259
		No	0.7703	
Z-Score		Yes	0.6667	0.7341
		No	0.7788	
Min-Max	60:40	Yes	0.6566	0.7208
		No	0.7648	
Z-Score		Yes	0.6682	0.7343
		No	0.7784	
Min-Max	65:35	Yes	0.6559	0.7226
		No	0.7676	
Z-Score		Yes	0.6711	0.7399
		No	0.7849	
Min-Max	70:30	Yes	0.6509	0.7172
		No	0.7623	
Z-Score		Yes	0.6830	0.7421
		No	0.7827	
Min-Max	75:25	Yes	0.6610	0.7242
		No	0.7675	
Z-Score		Yes	0.6758	0.7412
		No	0.7847	
Min-Max	80:20	Yes	0.6497	0.7181
		No	0.7641	
Z-Score		Yes	0.6662	0.7290
		No	0.7719	
Min-Max	85:15	Yes	0.6677	0.7278
		No	0.7695	
Z-Score		Yes	0.6677	0.7380
		No	0.7837	
Min-Max	**90:10**	Yes	0.6642	0.7242
		No	0.7660	
Z-Score		**Yes**	**0.6871**	**0.7496**
		No	**0.7914**	

stated that sometimes they would eat at the restaurant (65,374) and fast-food restaurant (67,543) before the COVID-19 pandemic, and most of them become never eat at the restaurant (102,845) and fast-food restaurant (104,831) after COVID-19 Pandemic.

For buying in markets or supermarkets or grocery stores, before COVID-19, most of the respondents said that they were often buying raw material for cooking and sometimes will buy ready-to-eat food products in these places, but after the COVID-19 pandemic, they will go these places to buy the items that they needed sometimes only. Therefore, we can conclude that the people reduce their spending and change of lifestyles will reduce the business transaction for the business owners, making their business drop and employees will lose the job due to the organisation being closed.

TABLE 7. Comparison of prediction models.

Prediction models	Split Data	Target	F1-score	Accuracy
Decision Tree	90:10	Yes	0.6324	0.6896
		No	0.7313	
Random Forest	90:10	Yes	0.7191	0.7545
		No	0.7820	
Naïve Bayes	90:10	Yes	0.6841	0.7215
		No	0.7510	
Gradient Boosted Tree	**80:20**	**Yes**	**0.6973**	**0.7589**
		No	**0.7997**	
Neural Network (MLP)	90:10	Yes	0.6871	0.7496
		No	0.7914	

B. MACHINE LEARNING MODELS

Based on the literature study, the five top performing machine learning algorithms were selected to predict financial vigilance of households. They are Decision Tree [20], [21], Random Forest [22], [23], Naïve Bayes [24], [25], Gradient Boosted Tree [23], [26], [27] and Neural Network [8], [28], [29]. In order to determine the optimal model for each of these algorithms, various parameter tuning was done and the best settings were identified. Later the optimal model from each algorithm was compared to identify the best machine learning model that can be further implemented to predict the households' financial vigilance. The target variable is financial preparedness, which is a yes/no question, with 55% saying yes and 45% saying no. Performance evaluation metrics F-measures and accuracy were calculated for all machine learning models and the best models were chosen based on F-score measure as it is the harmonic mean of precision and recall [26].

1) DECISION TREE MODEL

A decision tree algorithm is a tree-based machine learning that splits the data by sequence according to the target decision/variable [8]. Various parameter tuning was done to identify the optimal decision tree model, which is shown in Table 2. When 90% of the data is used for training the model and 10% is used for testing the model with gain ratio as the quality measure, the decision tree algorithm can predict households who are financially vigilant and not vigilant with F-score 0.6324 and 0.7313 respectively.

2) RANDOM FOREST MODEL

The random forest method is a machine learning technique that combines many decision trees and averages their output [21]. For example, Table 3 shows that when the type of partitioning is (90:10) with the information gain ratio criterion produces the best model.

TABLE 8. General profile of financially vigilant households.

Condition	Rul e Sup port	Rule confi dence
After COVID-19 Using taxi services / e-hailing services = "Never" AND Jobs = "Government Worker" AND Income Category = "T20 " AND If Income Changes, State the Reduction / Increase = "Not applicable" AND Know Financial Assistance to Employees? = "Yes" AND Sufficient Saving Period = "28-48"	93	100%
Work affected by the pandemic? = "I work from home" AND After COVID-19 Eat at the restaurant = "Never" AND Before COVID-19 Buy non-food products in malls and stores = "Sometimes" AND No. Household Members = "1-5" AND Income Changes due to pandemic? = "Remain" AND Income Category = "M40 " AND If Income Changes, State the Reduction / Increase = "Not applicable" AND Know Financial Assistance to Employees? = "Yes" AND Sufficient Saving Period = "28-48"	90	100%
Monthly Income = ">RM9000 " AND If Income Changes, State the Reduction / Increase = "Not applicable" AND Sufficient Saving Period = "52-96"	76	100%
Gender = "Male" AND Jobs = "Private Sector Employee" AND Income Category = "T20 " AND If Income Changes, State the Reduction / Increase = "Not applicable" AND Know Financial Assistance to Employees? = "Yes" AND Sufficient Saving Period = "28-48"	62	100%
Gender = "Male" AND Before COVID-19 Buy non-food products in malls and stores = "Often" AND No. Household Members = "1-5" AND Income Changes due to pandemic? = "Remain" AND Income Category = "M40 " AND If Income Changes, State the Reduction / Increase = "Not applicable" AND Know Financial Assistance to Employees? = "Yes" AND Sufficient Saving Period = "28-48"	58	100%
Main Activities Sector = "Industry: Supply of electricity, gas, steam and air conditioning" AND Income Changes due to Pandemic? = "Remain" AND If Income Changes, State the Reduction / Increase = "Not applicable" AND Sufficient Saving Period = "20-24"	57	100%
Sufficient Saving Period = "196-240" AND TRUE	54	100%

TABLE 9. General profile of households with lack of financial vigilant.

TABLE 9. *(Continued.)* General profile of households with lack of financial vigilant.

Condition	Rule Support	Rule confidence
Gender = "Male" AND Main Activities Sector = "No Information" AND Jobs = "Self-employed (no employees) : Running own business" AND Ethnic group = "Malay" AND Not working Because_EXTENT = "Working" AND Income Changes due to Pandemic? = "Reduction" AND Sufficient Saving Period = "<2"	130	99.23%
Before COVID-19 Using taxi services / e-hailing services = "Never" AND Jobs = "Self-employed (no employees) : Other Services Sector" AND Ethnic group = "Malay" AND Not working Because_EXTENT = "Working" AND Income Changes due to Pandemic? = "Reduction" AND Sufficient Saving Period = "<2"	122	100.00%
After COVID-19 Using taxi services / e-hailing services = "Never" AND Main Activities Sector = "No Information" AND Income Category = "B40" AND Jobs = "Private Sector Employee" AND Ethnic group = "Malay" AND Not working Because_EXTENT = "Working" AND Income Changes due to Pandemic? = "Reduction" AND Sufficient Saving Period = "<2"	110	95.45%
After COVID-19 Traveling for business / duty purposes = "Never" AND Ethnic group = "Chinese" AND Not working Because_EXTENT = "Working" AND Income Changes due to Pandemic? = "Reduction" AND Sufficient Saving Period = "<2"	94	98.94%
After COVID-19 Watching movies on the cinema = "Never" AND Main Activities Sector = "Services: Administration and support services" AND Jobs = "Government Worker" AND Ethnic group = "Malay" AND Not working Because_EXTENT = "Working" AND Income Changes due to pandemic? = "Reduction" AND Sufficient Saving Period = "<2"	87	100.00%
After COVID-19 Dining at Fast Food Restaurant = "Never" AND No. Employees In Company = "Not applicable" AND Jobs = "Self-employed (no employees): Food Preparation" AND Ethnic group = "Malay" AND Not working Because_EXTENT = "Working" AND Income Changes due to pandemic? = "Reduction" AND Sufficient Saving Period = "<2"	85	100.00%
Gender = "Male" AND Income Category = "M40 " AND Jobs = "Private Sector Employee" AND Ethnic group = "Malay" AND Not working Because_EXTENT = "Working" AND Income Changes due to Pandemic? = "Reduction" AND Sufficient Saving Period = "<2"	81	100.00%
Gender = "Female" AND Know Financial Assistance to Employees? = "No" AND Not working Because_EXTENT = "Non-Working: Still looking for a job" AND Income Changes due to pandemic? = "Non-Working" AND Sufficient Saving Period = "<2"	73	100.00%
No. Household Members = "1-5" AND Ethnic group = "Malay" AND Jobs = "Employer (have companies and employees)" AND Work affected by the pandemic? = "I lost my job because of Covid-19" AND Not working Because_EXTENT = "Working" AND Know Financial Assistance to Employees? = "Yes" AND Income Changes due to pandemic? = "Reduction" AND Sufficient Saving Period = "2-4"	73	100.00%
After COVID-19 Buy non-food products in malls and stores = "Never" AND Before COVID-19 Using taxi services / e-hailing services = "Never" AND After COVID-19 Eat at the restaurant = "Never" AND After COVID-19 Using taxi services / e-hailing services = "Never" AND After COVID-19 Watching movies on the cinema = "Never" AND Main Activities Sector = "Services: Other Services" AND Income Category = "B40" AND Jobs = "Private Sector Employee" AND Ethnic group = "Malay" AND Not working Because_EXTENT = "Working" AND Income Changes due to Pandemic? = "Reduction" AND Sufficient Saving Period = "<2"	66	100.00%

3) NAÏVE BAYES MODEL

Naïve Bayes algorithm applied Bayes Theorem using conditional probability to execute the prediction outcome [30] Table 4 shows the various parameter optimization of the

Naïve Bayes models and (90:10) partitioning is the optimal settings for this model.

4) GRADIENT BOOSTED TREE MODEL

The gradient boosted tree algorithm is similar to the random forest technique in that it averages all of the trees' initial results [21]. Table 5 shows the result of the gradient boosted tree algorithm. We found that the (80:20) type of partitioning was optimal in this model.

5) NEURAL NETWORK (MLP) MODEL

Multilayer Perceptron (MLP) is one of the neural network algorithms and it composed multiple perceptron to perform classification output [21]. There are two types of normalisation applied in the models: mix-max normalisation and z-score normalisation. After running all the results, we found that Z-score normalisation with (90:10) partitioning was optimal for this model. The result shows in Table 6.

Table 7 shows a comparison of optimal models from each algorithm. Overall, the target variable can be predicted with higher than 68.96 accuracy by all five machine learning algorithms. However, the gradient boosted tree algorithm is the best machine learning model because it can accurately predict households that are financially prepared and those that are not with 75.89 percent accuracy. Random forest is the second best method, followed by neural network (MLP) and naive bayes algorithms. Even though the data size is large, with 161,780 rows and 41 variables, all of these models perform best when 90% of the data is utilised to train the model and 10% to test the model.

Though most machine learning models do not explain the relationship between the predictors and the target variable, the decision tree algorithm can do so [31], [32]. Therefore, even though the decision tree did not provide the best model, we can examine the decision trees to understand the general profile of households with and without financial vigilance. Table 8 and Table 9 provide general profiles of households who are financially vigilant and those who are not. Rule support refers to the number of respondents who satisfy the conditions and rule confidence indicates correctness.

When characterizing and predicting financial status of households, machine learning models provide a higher overall model fit. This finding is supported by the existing study which recommended that when researchers and policymakers need to define and/or predict a household's future financial status, the machine learning approach can give a solid, efficient, and effective analytic method [28].

V. CONCLUSION

This study explored the COVID-19 survey dataset by leveraging data analytics and machine learning techniques to extract the insights. CRISP-DM method is followed and to create prediction models of household financial vigilance in dealing with the pandemic. In addition, five machine learning techniques were used: Decision Tree, Random Forest, Naive Bayes, Gradient Boosted Tree, and Multi-Layer Perceptron Neural Network. The best predictive model was found to be Gradient Boosted Tree.

The findings showed to what extent the pandemic has impacted households and provides machine learning models to identify the financially vigilant households. The B40 income group consists of 49% of the respondents, M40 13% and T20 3%. A total of 77% of these respondents revealed that their financial vigilance period was not more than eight weeks. Since this survey was conducted during the first wave of the pandemic, only 17% of the respondents lose their job.

However, the allocation of economic stimulus should not be based on the category of the household of B40 income group, ethnic group and age. The distribution of economic stimulus should target directly households in need of financial support based on their current employment status, number of members per household and location of residence. This knowledge could be useful to design suitable and effective economic stimulus packages.

REFERENCES

[1] L. L. Lim, "The socioeconomic impacts of COVID-19 in Malaysia: Policy review and guidance for protecting the most vulnerable and supporting enterprises," Int. Labour Org., Regional Office Asia Pacific, Bangkok, Thailand, Tech. Rep., 2020.

[2] W. Janssens, M. Pradhan, R. de Groot, E. Sidze, H. P. P. Donfouet, and A. Abajobir, "The short-term economic effects of COVID-19 on low-income households in rural Kenya: An analysis using weekly financial household data," *World Development*, vol. 138, pp. 1–8, Feb. 2021.

[3] A. U. M. Shah, S. N. A. Safri, R. Thevadas, N. K. Noordin, A. A. Rahman, Z. Sekawi, A. Ideris, and M. T. H. Sultan, "COVID-19 outbreak in Malaysia: Actions taken by the Malaysian government," *Int. J. Infectious Diseases*, vol. 97, pp. 108–116, Aug. 2020.

[4] R. G. Abdullah, N. I. Mersat, and S.-K. Wong, "Implications of COVID-19 pandemic on household food security: Experience from Sarawak, Malaysia," *Int. J. Bus. Soc.*, vol. 22, no. 1, pp. 1–13, Mar. 2021.

[5] *Prihatin Rakyat Economic Stimulus Package*, Prime Minister Office, Putrajaya, Malaysia, 2020.

[6] A. Hayes. (2020). *Stimulus Package Definition*. [Online]. Available: https://investopedia.com

[7] S. Flanders, M. Nungsari, and H. Y. Chuah, "The COVID-19 hardship survey: An evaluation of the Prihatin Rakyat economic stimulus package," Asia School Bus. Discuss. Paper Ser., Asia School Bus. COVID-19 Hardship Surv., Eval. Prihatin Rakyat Econ. Stimulus Package, Asia School Bus., Kuala Lumpur, Malaysia, Tech. Rep., Apr. 2020, pp. 1–44.

[8] N. Sangavi, R. Jeevitha, P. Kathirvel, and K. Premalatha, "Impact of classification algorithms on census dataset," *Int. J. Recent Technol. Eng.*, vol. 8, no. 5, pp. 2666–2670, 2020, doi: 10.35940/ijrte.e6027.018520.

[9] J. Tang and T. Begazo. (2020). *Digital Stimulus Packages: Lessons Learned and What's Next*. Accessed: May 15, 2021. [Online]. Available: https://blogs.worldbank.org/digital-development/digital-stimulus-packages-lessons-learned-and-whats-next

[10] N. S. Alkhatri, N. Zaki, E. Mohammed, and M. Shallal, "The use of data mining techniques to predict the ranking of e-Government services," in *Proc. 12th Int. Conf. Innov. Inf. Technol. (IIT)*, Nov. 2016, pp. 1–6.

[11] J. Lu, "Data analytics research-informed teaching in a digital technologies curriculum," *Informs Trans. Educ.*, vol. 20, no. 2, pp. 57–72, Jan. 2020.

[12] C. Alexopoulos, Z. Lachana, A. Androutsopoulou, V. Diamantopoulou, Y. Charalabidis, and M. A. Loutsaris, "How machine learning is changing e-Government," in *Proc. 12th Int. Conf. Theory Pract. Electron. Governance*, Apr. 2019, pp. 354–363.

[13] L. Li, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, K. Cao, D. Liu, G. Wang, Q. Xu, X. Fang, S. Zhang, J. Xia, and J. Xia, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, Mar. 2020, Art. no. 200905.

[14] M. M. Mostafa and A. A. El-Masry, "Citizens as consumers: Profiling e-government services' users in Egypt via data mining techniques," *Int. J. Inf. Manage.*, vol. 33, no. 4, pp. 627–641, Aug. 2013.

[15] Y. Chen, J. Wang, and Z. Cai, "Study on the application of machine learning in government service: Take consumer protection service as an example," in *Proc. 15th Int. Conf. Service Syst. Service Manage. (ICSSSM)*, Jul. 2018, pp. 1–5.

[16] E. Loukis, N. Kyriakou, and M. Maragoudakis, "Using government data and machine learning for predicting firms' vulnerability to economic crisis," in *Proc. Int. Conf. Electron. Government*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 12219, 2020, pp. 345–358.

[17] A. Tiwari, A. V. Dadhania, V. A. B. Ragunathrao, and E. R. A. Oliveira, "Using machine learning to develop a novel COVID-19 vulnerability index (C19VI)," *Sci. Total Environ.*, vol. 773, Jun. 2021, Art. no. 145650.

[18] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Proc. Comput. Sci.*, vol. 181, pp. 526–534, Jan. 2021.

[19] S. Tangirala, "Evaluating the impact of Gini index and information gain on classification using decision tree classifier algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, pp. 612–619, 2020.

[20] C. Zhang, C. Hu, S. Xie, and S. Cao, "Research on the application of decision tree and random forest algorithm in the main transformer fault evaluation," *J. Phys., Conf. Ser.*, vol. 1732, no. 1, Jan. 2021, Art. no. 012086.

[21] N. Sangavi, R. Jeevitha, P. Kathirvel, and K. Premalatha, "Impact of classification algorithms on census dataset," *Int. J. Recent Technol. Eng.*, vol. 8, no. 5, pp. 2666–2670, 2020.

[22] C. M. Liyew and H. A. Melese, "Machine learning techniques to predict daily rainfall amount," *J. Big Data*, vol. 8, no. 1, pp. 1–11, Dec. 2021.

[23] H. Qian, B. Wang, M. Yuan, S. Gao, and Y. Song, "Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree," *Expert Syst. Appl.*, vol. 190, Mar. 2022, Art. no. 116202.

[24] J. F. Easton, H. R. Sicilia, and C. R. Stephens, "Classification of diagnostic subcategories for obesity and diabetes based on eating patterns," *Nutrition Dietetics*, vol. 76, no. 1, pp. 104–109, Feb. 2019.

[25] S. K. Trivedi, "A study on credit scoring modeling with different feature selection and machine learning approaches," *Technol. Soc.*, vol. 63, Nov. 2020, Art. no. 101413.

[26] R. Kannan, I. Z. W. Wang, H. B. Ong, K. Ramakrishnan, and A. Alamsyah, "COVID-19 impact: Customised economic stimulus package recommender system using machine learning techniques," *FResearch*, vol. 10, p. 932, Nov. 2021.

[27] Z. Qin, L. Yan, H. Zhuang, Y. Tay, R. K. Pasumarthi, X. Wang, M. Bendersky, and M. Najork, "Are neural rankers still outperformed by gradient boosted decision trees?" in *Proc. 9th Int. Conf. Learn. Represent.*, May 2021. [Online]. Available: https://openreview.net/forum?id=Ut1vF_q_vC

[28] W. Heo, J. M. Lee, N. Park, and J. E. Grable, "Using artificial neural network techniques to improve the description and prediction of household financial ratios," *J. Behav. Exp. Finance*, vol. 25, Mar. 2020, Art. no. 100273.

[29] K. Lepenioti, M. Pertselakis, A. Bousdekis, A. Louca, F. Lampathaki, D. Apostolou, G. Mentzas, and S. Anastasiou, "Machine learning for predictive and prescriptive analytics of operational data in smart manufacturing," in *Proc. Int. Conf. Adv. Inf. Syst. Eng.*, in Lecture Notes in Business Information Processing, vol. 382, 2020, pp. 5–16.

[30] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective Naïve Bayes algorithm," *Knowl.-Based Syst.*, vol. 192, Mar. 2020, Art. no. 105361.

[31] J. D. Kelleher, B. M. Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, 1st ed. Cambridge, MA, USA: MIT Press, 2015.

[32] T. Hastie, R. Tibshirani, G. James, and D. Witten, *An Introduction to Statistical Learning* (Springer Texts in Statistics), vol. 102, 2nd ed. New York, NY, USA: Springer, 2021, p. 618.

RATHIMALA KANNAN (Senior Member, IEEE) received the B.Sc. degree in applied sciences-computer technology and the Master of Computer Applications (M.C.A.) degree from Bharathiar University, Coimbatore, India, and the Ph.D. degree from Multimedia University, Malaysia. Her Ph.D. research was in the area of web structure mining and titled as Characterization of E-Commerce Website Structures Using Webometrics and Social Network Analysis Methods. She is currently a Senior Lecturer with the Department of Information Technology, Faculty of Management, Multimedia University. She has published in the *International Journal of Electronic Business* and *Journal of Theoretical and Applied Electronic Commerce Research*. She has also worked as an industrial consultant in developing E-Commerce portal. Her research interests include business data analytics, healthcare analytics, data mining for business intelligence, and information systems. She is a Senior Member of IEEE Computer Society, IEEE Malaysia Consultants Network Affinity Group, and the International Association of Computer Science and Information Technology. She is also a Professional Technologist Member in Malaysian Board of Technologist (MBOT). She is a HRDF Certified Trainer.

KHOR WOON SHING received the Diploma degree (Hons.) in business management from the Sultan Abdul Samad Vocational College and the B.A. degree (Hons.) in enterprise management system from Multimedia University, Malaysia. Her B.A. degree final year project research was in area of data analytics and titled "Leveraging Machine Learning to Understand Effect of Covid-19 on the Economy and Individual." Her research interests include the impact of Covid-19 to the economy and individual, application of machine learning in building prediction models, and the uses of machine learning in segmentation of citizens.

KANNAN RAMAKRISHNAN (Senior Member, IEEE) received the Bachelor of Engineering degree in electronics and communication and the Postgraduate Diploma degree in medical instrumentation technology from the Coimbatore Institute of Technology, India, the Master of Engineering degree in computer science from the Regional Engineering College (now known as the National Institute of Technology), Trichy, India, and the Ph.D. degree in information technology from Multimedia University, Malaysia. He is currently working with the Faculty of Computing and Informatics, Multimedia University. He has got the teaching and research experience of over 30 years. He has given keynote talks, served as a technical and advisory committee member for international conferences, served as a Ph.D. thesis examiner for different universities, and led a group of projects funded by Malaysian Government, Telekom Malaysia, and Intel. His current research interests include biomedical data processing, machine learning, and cybersecurity. He is a Life Member of the Indian Society for Technical Education, a Senior Member of the IEEE Signal Processing Society and the IEEE Engineering in Medicine and Biology Society, and a member of Technical Committee on Health Informatics Standards, Standards Malaysia.

HWAY BOON ONG received the M.Sc. degree in economics from University Putra Malaysia, the M.Phil. degree in banking from MMU, and the Ph.D. degree in financial economics from University Putra Malaysia. She is currently working as an Associate Professor at Multimedia University, Cyberjaya Campus. She has published in several internationally peer-reviewed journal articles. Her research interests include money, banking, and social economics.

ANDRY ALAMSYAH (Member, IEEE) received the B.S. degree in mathematics from ITB Bandung, in 1996, the M.S. degree in information system from the Universite de Picardie Jules Verne, France, in 2004, and the Ph.D. degree in electrical engineering and informatics from ITB Bandung, in 2017.

Since 2011, he has been a Researcher at the School of Economics and Business, Telkom University. He was the Director of the Digital Business Ecosystem Research Center, from 2018 to 2020. He is currently the Founder and the Chief of the Laboratory Social Computing and Big Data. He is also the Founder and the President of the Asosiasi Ilmuwan Data Indonesia (AIDI), the only formal association for data scientists in Indonesia. By 2021, he has appointed as an Honorary Member of Asosiasi Blockchain Indonesia. His research interests include social computing, social networks, complex networks/network science, big data, blockchain technology and opportunities, technopreneurship, and business model.