Check for updates

RESEARCH ARTICLE

REVISED COVID-19 impact: Customised economic stimulus package recommender system using machine learning techniques [version 2; peer review: 1 approved, 2 approved with reservations]

Rathimala Kannan[1], Ivan Zhi Wei Wang[2], Hway Boon Ong[3], Kannan Ramakrishnan[4], Andry Alamsyah[5]

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

v2 **First published:** 16 Sep 2021, **10**:932
https://doi.org/10.12688/f1000research.72976.1
Latest published: 12 Nov 2021, **10**:932
https://doi.org/10.12688/f1000research.72976.2

Abstract

Background: The Malaysian government reacted to the pandemic's economic effect with the Prihatin Rakyat Economic Stimulus Package (ESP) to cushion the novel coronavirus 2019 (COVID-19) impact on households. The ESP consists of cash assistance, utility discount, moratorium, Employee Provident Fund (EPF) cash withdrawals, credit guarantee scheme and wage subsidies. A survey carried out by the Department of Statistics Malaysia (DOSM) shows that households prefer different types of financial assistance. These preferences forge the need to effectively customise ESPs to manage the economic burden among low-income households. In this study, a recommender system for such ESPs was designed by leveraging data analytics and machine learning techniques.

Methods: This study used a dataset from DOSM titled "Effects of COVID-19 on the Economy and Individual - Round 2," collected from April 10 to April 24, 2020. Cross-Industry Standard Process for Data Mining was followed to develop machine learning models to classify ESP receivers according to their preferred subsidies types. Four machine learning techniques—Decision Tree, Gradient Boosted Tree, Random Forest and Naïve Bayes—were used to build the predictive models for each moratorium, utility discount and EPF and Private Remuneration Scheme (PRS) cash withdrawals subsidies. The best predictive model was selected based on F-score metrics.

Results: Among the four machine learning techniques, Gradient Boosted Tree outperformed the rest. This technique predicted the

Open Peer Review

Reviewer Status ? ? ✓

	Invited Reviewers		
	1	**2**	**3**
version 2 (revision) 12 Nov 2021			✓ report
version 1 16 Sep 2021	? report	? report	? report

1. **Setia Pramana**, Politeknik, Statistika STIS, Jakarta, Indonesia
BPS Statistics Indonesia, Jakarta, Indonesia

2. **Shahrinaz Ismail**, Albukhary International University, Alor Setar, Malaysia

3. **Sreeja N.K**, PSG College of Technology, Coimbatore, India

Any reports and responses or comments on the article can be found at the end of the article.

following: moratorium preferences with 93.8% sensitivity, 82.1% precision and 87.6% F-score; utilities discount with 86% sensitivity, 82.1% precision and 84% F-score; and EPF and PRS with 83.6% sensitivity, 81.2% precision and 82.4% F-score. Households that prefer moratorium subsidies did not favour other financial aids except for cash assistance.
Conclusion: Findings present machine learning models that can predict individual household preferences from ESP. These models can be used to design customised ESPs that can effectively manage the financial burden of low-income households.

Keywords
COVID-19, low-income households, economic stimulus package, customisation, data analytics, machine learning, Gradient Boosted Tree

This article is included in the Research Synergy Foundation gateway.

Corresponding author: Rathimala Kannan (rathimala.kannan@mmu.edu.my)

Author roles: Kannan R: Conceptualization, Writing – Original Draft Preparation; **Wang IZW**: Data Curation; **Ong HB**: Writing – Review & Editing; **Ramakrishnan K**: Writing – Review & Editing; **Alamsyah A**: Validation

Competing interests: No competing interests were disclosed.

Grant information: This work was partly supported by Multimedia University.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Kannan R *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Kannan R, Wang IZW, Ong HB *et al.* **COVID-19 impact: Customised economic stimulus package recommender system using machine learning techniques [version 2; peer review: 1 approved, 2 approved with reservations]** F1000Research 2021, **10**:932 https://doi.org/10.12688/f1000research.72976.2

First published: 16 Sep 2021, **10**:932 https://doi.org/10.12688/f1000research.72976.1

REVISED **Amendments from Version 1**

The term "How" was added to the research questions. The methodology section has been enhanced to include different of machine learning approaches. A new flowchart (new Figure 2) of the modelling and evaluation process has been introduced, which demonstrates how a supervised machine learning model is constructed, as well as parameter tuning to achieve optimised models. To select the optimal model, k-fold cross validation was used, which revealed no differences when the data was separated into training and testing datasets. The discussion section has been updated to address each research questions individually, providing additional detail on the findings and their implications. The trade-off between prediction accuracy and interpretability of a model in machine learning approach is explained. The part on limitations has been explained in detail.

Any further responses from the reviewers can be found at the end of the article

Introduction

The novel coronavirus 2019 (COVID-19) pandemic has created devastation in people's lives worldwide, both socially and economically (Shah *et al.*, 2020). As a result, governments have adopted various strategies aimed at reducing the pandemic's impact, particularly the financial strain. The Malaysian government has introduced a series of economic stimulus packages to support various segments of its citizens. One such support is the Prihatin Rakyat Economic Stimulus Package (ESP) to cushion the impact of COVID-19 on low-income households after the first movement control in the country. The ESP consists of cash assistance, utility discount, moratorium, Employee Provident Fund and Private Remuneration Scheme (EPF and PRS) cash withdrawals and Credit Guarantee Scheme and Wage subsidies (Flanders *et al.*, 2020). Following the implementation of ESP, the Department of Statistics Malaysia (DOSM) carried out a special survey from April 10 to April 24,2020 to better understand the implications of COVID-19 on the economy and households. The study included questions on social and economic factors and subsidy preferences.

A typical low-income household often bears considerable debt and has limited savings. When movement control was implemented, households that lost their income sources faced difficulties in accessing necessities, such as food and housing (Flanders *et al.*, 2020). Even though the government offered ESP to help residents cope financially, the demands and desires of citizens in the event of a pandemic are unknown. For example, several households are reluctant to withdraw from EPF and PRS due to its reduction on their savings for old age. A personalised ESP can be built to reduce residents' financial burden in this crisis if we can foresee their requirements and preferences for various subsidies, such as cash allowance, utility discount, moratorium or EPF and PRS withdrawals. Using data analytics and machine learning approaches, this study attempted to analyse survey data and construct predictive models for customised economic stimulus packages. The following research questions were put forward.

1. How can households that favour moratorium subsidies be identified?

2. How to find out which households seek utility discount subsidies?

3. How can households who desire EPF and PRS withdrawal subsidies be identified?

This study contributes to the literature by using four machine learning techniques on socioeconomic survey data and predicting household subsidy preferences. A comparison of the feature selection methods, such as Gini index, Gain–Ratio and various partitioning ratios of the training and test data sets were carried out. The outcomes of this study can help the government deliver better and improved stimulus packages in the future based on individual preferences.

Methods

For planning and execution, this study used the Cross-Industry Standard Process-Data Mining (CRISP-DM), which is the industry-independent *de-facto* standard for implementing data mining initiatives (Schröer *et al.*, 2021). This process has six phases, namely, business understanding, data understanding, data preparation, modelling, evaluation and deployment. Figure 1 depicts the activities carried out in each phase, as further explained below.

Business understanding

This phase identifies the problems to solve using the machine learning perspective and approach. All three research questions were selected as the problems, and the purpose was to propose predictive models for the moratorium, utility discount and EPF and PRS subsidies in the Prihatin Rakyat ESPs.

Data understanding

Data gathering, evaluating, characterising and assuring its quality are part of this phase. DOSM performed a special survey (Round 2) to investigate the consequences of the COVID-19 epidemic on household economics and status ("Department of Statistics Malaysia Official Portal," 2020; Malaysia, 2020). The dataset includes 36 questions and 41,386 respondents. However, the data obtained from DOSM were not complete due to missing questions. The missing questions were Q3, Q6, Q19, and Q27 – Q31. In terms of the total respondents, the data were complete and had a total of 41,386 participants, all of them were aged 15 and older. 96.8% of the respondents have received benefits from Prihatin Rakyat ESPs. The raw data were based on responses from respondents, which included qualitative personal opinions on economy, employment, lifestyle, and education. The original dataset was in Malay language, and is translated into English for this study and given below as Table 1.

Data preparation

There were 28 questions available for further analysis in this study, eliminating the missing ones. Question 32 and 33 were

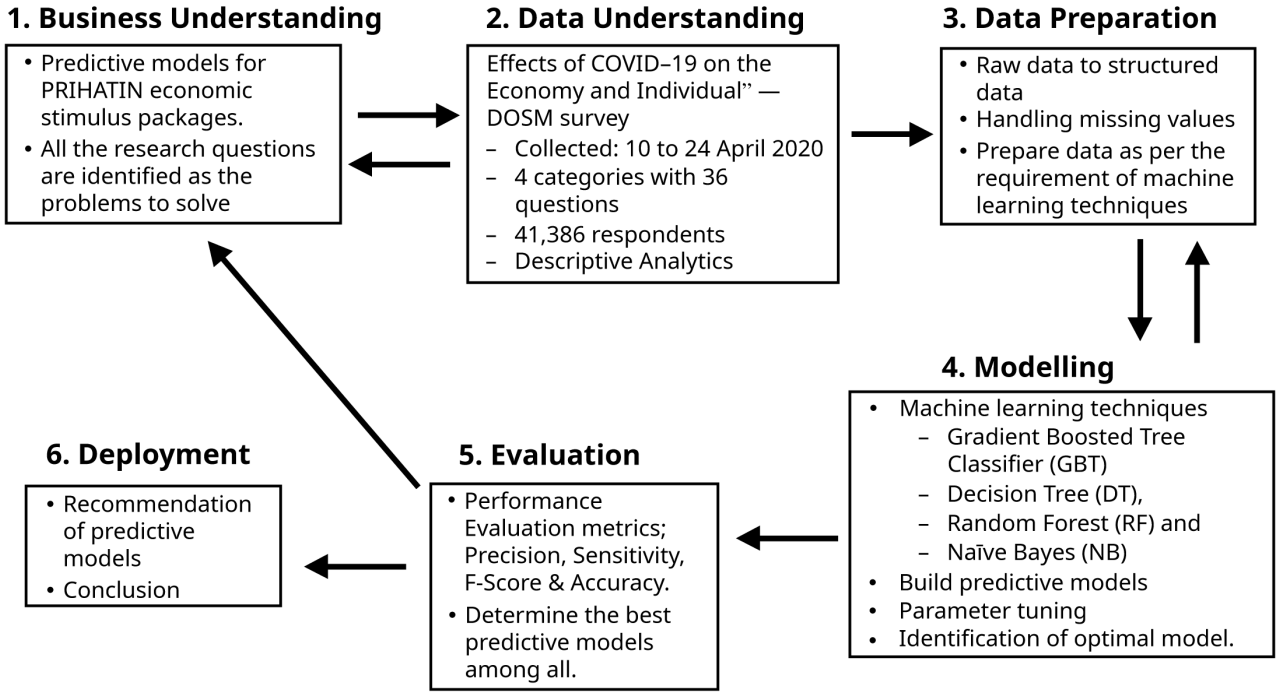

Figure 1. Research methods.

excluded from the survey because they focused on the primary food and non-food products purchased during the time of movement control orders. Given that the dataset was cluttered with missing values and errors, a considerable effort was spent on its cleaning before applying descriptive analytics techniques. Q34, Q35 and Q36 had missing values of 2071, 2243 and 2310 respectively and were replaced by the most frequent values. By cleaning the data, the raw data is transformed into structured data. Without losing any information, all lengthy responses were reduced to short and detailed responses. If the original answer for respondent's dwelling state was "Wilayah Persekutuan Kuala Lumpur," it was converted to "KL." Questions with answers were labelled "Yes," whereas those without answers are labelled "No." One question, for example, inquired about respondents' willingness to eat out as part of the new norm's lifestyle adjustments. Those who agreed said they "will not eat out." Those who opposed to the shift in lifestyle left the question unanswered. As a result, these questions were changed to a "Yes" or "No" format.

Modelling

This phase employed a variety of machine learning techniques in order to achieve the study's goal of developing classification models. Different approaches, such as information-based learning, similarity-based learning, error-based learning, and probability-based learning, can be used to create classifiers (Kelleher *et al.*, 2015). This research used information based learning (a decision tree algorithm family) as the best model for explaining decision logic. To create prediction models, we used four machine learning techniques: Decision Tree, Random Forest, Gradient Boosted Tree and Naïve Bayes. These four machine learning techniques were chosen from a literature review (Mostafa *et al.*, 2021; Sangavi *et al.*, 2020) and used to determine the optimal model by tuning their parameters. Feature selection methods, such as Gini index, Gain–Ratio and various partitioning ratios of the training and test data sets were also compared (Trivedi, 2020). Figure 2 depicts parameter tuning carried out in the modelling phase.

Evaluation

In this phase, the best predictive model for each subsidy was selected based on the standard performance evaluation metrics: Sensitivity, Precision, F-Score and Accuracy (Moscato *et al.*, 2021). The formulas used to calculate each of the metrics are given below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

The completeness of a prediction model was measured by sensitivity, also known as recall and true positive rate (TPR). This metric determined the proportion of positive predictions by a model that corresponds to true positive values (Moscato *et al.*, 2021). The formula is given below.

$$Sensitivity = \frac{TP}{TP+FN}$$

Precision in data analytics refers to a model's ability to correctly forecast outcomes. In other words, precision is a true positive divided by a combination of true and false positives.

Table 1. Survey questions.

1. Joined and answered Survey Round 1 (Yes, No)
2. State (KL, Johor, Selangor, Perak, Sarawak, Kedah, Pahang, Putrajaya, Perlis, Kelantan, Melaka, Pulau Pinang, Labuan, Terengganu, Negeri Sembilan, Sabah)
3.
4. Gender (Female, Male)
5. Stay in (City, Rural)
6.
7. Age group (15–24, 25–34, 35–44, 45–54, 55–64, 65 and above)
8. Marital status
9. (Single/Non-married, Married, Single mother, Widow/Widower, Divorced/Separated)
10. Number of dependents (including respondents) (1–2, 3–4, 5–10, More than 10 people)
11. Ethnic groups (Malay, Indian, Native Sabah/Sarawak, Chinese, Others, Foreigner)
12. Benefit from the PRIHATIN Economic Stimulus Package (ESP)? (Yes, No)
13. The most advantageous form of assistance received under the PRIHATIN Economic Stimulus Package?
    a. Cash assistance (National Prihatian assistance, IPT student assistance, E-hailing) (Yes, No)
    b. Utilities discount (Yes, No)
    c. Moratorium (Yes, No)
    d. EPF cash withdrawals & Private remuneration scheme (Yes, No)
    e. Credit guarantee scheme (Yes, No)
    f. Wage subsidies and payments under ERP program (Yes, No)
    g. NA (Yes, No)
13. Process/procedure achievement level for PRIHATIN assistance (Easy, Difficult, Medium, NA, Others)
14. Satisfied with the help of PRIHATIN (Yes, No, NA)
15. Status of eligibility as a recipient of PRIHATIN aid (Eligible, New application, Appeal, Not eligible)
16. Ranking according to your preferences with the situation facing now [Health & lives] (1,2,3)
    a. Ranking according to your preferences with the current situation [Income (employment/ business/enterprise)] (1,2,3)
    b. Ranking according to your preferences with the situation faced now [Life goes back as normal] (1,2,3)
17. The impact of the PRIHATIN Economic Stimulus Package (Effective, Most effective, Others, NA, No effects)
18. With the extension of the Movement Control Order (MCO) until 28 April 2020, does it still require assistance/ the PRIHATIN Economic Stimulus Package for the next phase? (Yes, No)
19.
20. Views on health workers (frontlines) and health facilities provided by the government in dealing with COVID-19 (Good, Bad)
21. Views on achievements and facilities provided in dealing with COVID-19 in Malaysia compared to other countries. (Good, Bad)
22. COVID-19 outbreak affects lifestyle (Yes, No)
23. Ready to have a new normal lifestyle (i.e. a new normal life will differ from normal life before the spread of the COVID-19 outbreak) (Yes, No)
24. If ready, lifestyle changes will be done
    a. Number of lifestyle changes to be done (Yes, No)
    b. Will not eat out (Yes, No)
    c. Limiting social activities (Yes, No)
    d. Limiting sports and recreational activities (Yes, No)
    e. Limited religious and spiritual activity at home (Yes, No)
    f. Limiting tourism activities (Yes, No)
    g. Improves hygiene (Yes, No)
    h. Others (Yes, No)
    i. NA (Yes, No)
25. Working during the MCO period (Work from home, Not working, Rotation - partial payment, Full paid leave, Rotation - full payment, Semi-salary leave, Retrenched)
26. Work as (Government servants, Private sector employees, Not working, Self-employed, GLC employees, MNC employees, employers, Unpaid family workers)
27.
28.
29.
30.
31.
32. Main expenses on food products during MCO (Yes, No)
    a. Number of major food products expenditure during MCO
    b. Dry food items (e.g. bihun, biscuits, nestum, etc.)
    c. Cooked food (takeaway/delivery)
    d. Instant noodles
    e. Eggs
    f. Fish/chicken/meat/seafood for cooking
    g. Vegetables
    h. Fruits
    i. Frozen food (e.g. sausages, nuggets, fish balls, french fries, etc.)
    j. Rice
    k. Spaghetti
    l. Bread
    m. Baby food
    n. Animal food
    o. Cooking oil
    p. Essential items for cooking (example: shallots, garlic, sugar, salt, etc.)
    q. Canned beverages (examples: milo, condensed milk, powdered milk, etc.)
    r. Flour (including wheat flour, Rice flour, glutinous flour, etc.)
    s. 3-in-1 Drinks
    t. Vitamins/Supplements
    u. Drinking water
    v. Othersr
33. Main expenses for non-food products during MCO (Yes, No)
    a. Number of main expenses on non-food during MCO
    b. Hand wash
    c. Tissue paper
    d. Disinfectant
    e. Face mask
    f. Baby diaper
    g. Wet tissue
    h. Sanitary pad
    i. Laundry soap
    j. Toiletries
    k. Thermometer
    i. Medications (e.g. fever medicine, flu medicine, cough medicine, etc.)
    m. First aid kit
    n. Gloves (various sizes)
    o. Others
34. Internet access level for yourself / your child's online learning (NA, Slow, Fast, Average, No Internet access)
35. Average hours per day of yourself/child online learning during the MCO period (NA, <1, 1–3, 3–5, 5–8, >8)
36. Accessibility of faster Internet speed during online learning (NA, Morning, Night, Afternoon)

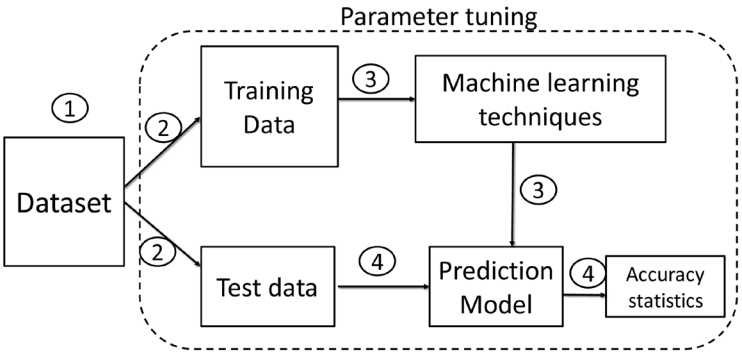

Figure 2. Flowchart for modelling and evolution process.

$$Precision = \frac{TP}{TP+FP}$$

F-Score, also known as F1 Score, is a balance of both precision and sensitivity. Hence, this study used F-Score to evaluate the machine learning models.

$$F\ Score = 2\times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

Deployment

In the final phase a deployment strategy for the model was created and documented. The best predictive model as determined for each of the subsidies was to be recommended for further deployment. The entire CRISP-DM phases were carried out using the Konstanz Information Miner (KNIME 4.3.2), a free and open-source data analytics software.

Ethical approval

This study was carried out from November 23, 2020 to October 06, 2021 and has obtained ethics approval (EA1322021) by Technology Transfer Office, secretariat of research ethics committee, Multimedia university.

Results

The outcomes of this study were organised as descriptive analytics, model optimisation and findings. Descriptive analytics helps to understand the characteristics of each respondent and the relationship between variables. Table 2 provides the descriptive information on the respondents.

Figure 3 shows the various types of subsidies offered in the ESP. Among the 41,386 respondents, 72.2% were eligible to receive subsidies, 21.9% were newly applied, 3.2% were not eligible and 2.7% had appealed. Figure 3 also shows the most beneficial forms of support. The most popular type of subsidy was cash allowance, followed by moratorium, utility discounts and EPF and PRS cash withdrawals. The least preferred type was the credit guarantee plan and wage subsidies.

Following the descriptive analytics, the four machine learning techniques were applied to develop prediction models for each moratorium, utility discount and EPF withdrawals subsidies. Decision Tree, Gradient Boosted Tree, Random Forest and

Table 2. Descriptive statistics of respondent demographics.

State		Age Group	
Selangor	27.74%	35–44	38.87%
Johor	12.35%	25–34	29.19%
Sabah	8.41%	45–54	19.38%
KL	7.84%	15–24	6.6%
Perak	6.29%	55–64	5.46%
Kedah	4.96%	65 and above	0.50%
Pahang	4.87%	**Marital Status**	
Melaka	4.83%	Married	70.95%
Sarawak	4.52%	Single/ Non-Married	24.35%
Negeri Sembilan	4.38%	Divorced/ Separated	1.98%
Kelantan	3.30%	Single mother	1.42%
Pulau Pinang	3.23%	Widow/ Widower	1.28%
Terengganu	3.02%	**Location**	
Putrajaya	2.94%	City	70.71%
Perlis	0.83%	Rural	29.28%
Labuan	0.50%		
Gender		**Number of dependents including self**	
Female	55.7%	3–4	36.21%
Male	44.3%	5–10	35.2%
		1–2	28.1%
		More than 10 people	0.47%
Ethnic Group		**Job status**	
Malay	79.61%	Government Servants	39.22%
Native Sabah/ Sarawak	10.29%	Private Sector Employees	27.14% 21.61%
Chinese	7.18%	Not Working	4.35%
Indian	1.75	GLC Employees	3.05%
Others	1.07	Self-Employed	2.18%
Foreigner	0.1%	Employers	2.12%
		MNC Employees	0.32%
		Unpaid family workers	

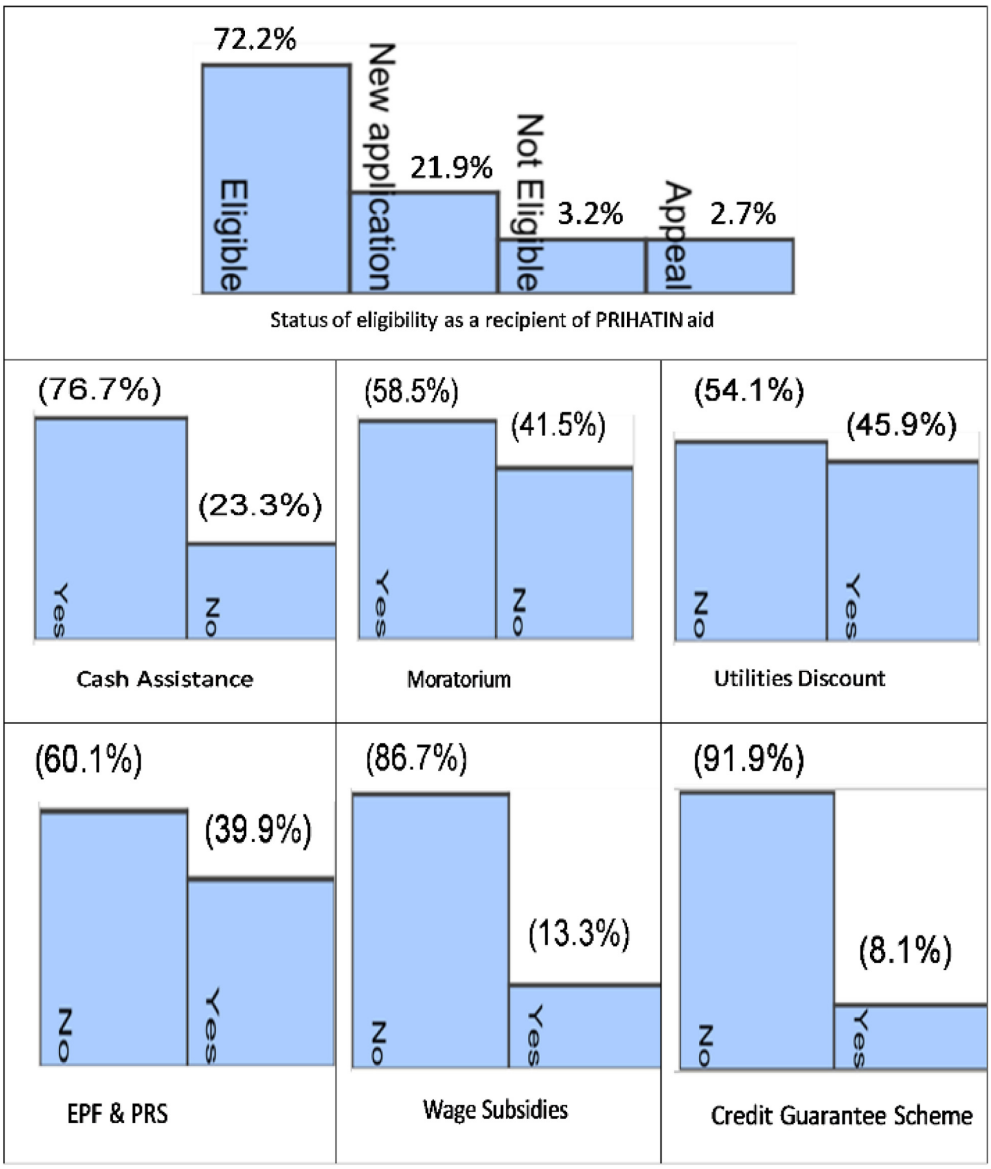

Figure 3. Most beneficial forms of assistance received under the Prihatin Rakyat ESP.

Naïve Bayes are subjected to parameter tuning to determine the best model and parameter values.

Table 3 to Table 6 show how the optimal model was obtained from each machine learning technique. Partitioning ratio indicates the training and test data. Gain ratio, Gini index and information gain were used to measure the quality of each predictor in classifying the target variable. The results show that the Gradient Boosted Tree and Naïve Bayes techniques performed well when 60% of the data were used to train the machine learning models and the other 40% was used for testing. Random Forest and Decision Tree techniques generated the best models when the training data were 80% and the test data were 20%. F-Score was used as the evaluation measure to select the optimal models. After identifying the optimal models, the best was selected among the four machine learning techniques. Table 7 shows the results. Gradient Boosted Tree outperformed the rest of the techniques in predicting the moratorium preference with 93.8% sensitivity, 82.1% precision and 87.6% F-score. When the data is partitioned with k= 5, and K=10, the findings reveal little difference in classifier performance, and the Gradient boosting tree still performs the best. When the

Table 3. **Gradient Boosted Tree - Parameter tuning and identification of optimal model.**

Machine Learning Technique	Partitioning Ratio	Target Variable: Moratorium	Precision	Sensitivity	F-score	Accuracy
Gradient Boosted Tree	50:50	No	0.877	0.71	0.784	0.838
		Yes	0.819	0.929	0.871	
	60:40	No	0.891	0.711	**0.791**	**0.844**
		Yes	0.821	**0.938**	**0.876**	
	70:30	No	0.883	0.717	0.791	0.843
		Yes	0.823	0.933	0.874	
	80:20	No	0.881	0.701	0.781	0.837
		Yes	0.815	0.933	0.87	

Table 4. **Naïve Bayes - Parameter tuning and identification of optimal model.**

Machine Learning Technique	Partitioning Ratio	Target Variable: Moratorium	Precision	Sensitivity	F-score	Accuracy
Naïve Bayes	50:50	No	0.627	0.449	0.524	0.661
		Yes	0.675	0.811	0.737	
	60:40	No	0.64	0.453	**0.53**	0.667
		Yes	0.679	0.819	**0.742**	
	70:30	No	0.633	0.449	0.526	0.663
		Yes	0.676	0.815	0.739	
	80:20	No	0.624	0.443	0.518	0.658
		Yes	0.673	0.811	0.735	

Table 5. **Decision Tree - Parameter tuning and identification of optimal model.**

Machine Learning Technique	Partitioning Ratio	Target Variable: Moratorium	Precision	Sensitivity	F-score	Accuracy
Decision Tree	50:50 (Gini Index)	No	0.776	0.705	0.739	0.793
		Yes	0.804	0.856	0.829	
	50:50 (Gain Ratio)	No	0.745	0.685	0.731	0.772
		Yes	0.789	0.834	0.811	
	60:40 (Gini Index)	No	0.772	0.697	0.733	0.789
		Yes	0.799	0.855	0.826	
	60:40 (Gain Ratio)	No	0.746	0.669	0.705	0.768
		Yes	0.781	0.839	0.809	
	70:30 (Gini Index)	No	0.766	0.7	0.731	0.787
		Yes	0.8	0.848	0.823	
	70:30 (Gain Ratio)	No	0.75	0.685	0.716	0.775
		Yes	0.79	0.838	0.813	
	80:20 (Gini Index)	No	0.779	0.727	**0.752**	**0.802**
		Yes	0.816	0.854	**0.834**	
	80:20 (Gain Ratio)	No	0.743	0.697	0.719	0.774
		Yes	0.794	0.83	0.812	

Table 6. **Random Forest - Parameter tuning and identification of optimal model.**

Machine Learning Technique	Partitioning Ratio	Target Variable: Moratorium	Precision	Sensitivity	F-score	Accuracy
Random Forest	50:50 (Information Gain)	No	0.809	0.766	0.787	0.828
		Yes	0.84	0.872	0.856	
	50:50 (Gain Ratio)	No	0.805	0.744	0.773	0.819
		Yes	0.828	0.872	0.849	
	50:50 (Gini Index)	No	0.804	0.763	0.783	0.825
		Yes	0.838	0.868	0.853	
	60:40 (Information Gain)	No	0.809	0.752	0.779	0.823
		Yes	0.833	0.874	0.853	
	60:40 (Gain Ratio)	No	0.804	0.739	0.77	0.817
		Yes	0.825	0.873	0.848	
	60:40 (Gini Index)	No	0.803	0.759	0.78	0.823
		Yes	0.836	0.868	0.852	
	70:30 (Information Gain)	No	0.808	0.774	0.791	0.83
		Yes	0.845	0.87	0.857	
	70:30 (Gain Ratio)	No	0.821	0.753	0.786	0.83
		Yes	0.835	0.884	0.859	
	70:30 (Gini Index)	No	0.815	0.765	0.789	0.831
		Yes	0.84	0.877	0.858	
	80:20 (Information Gain)	No	0.813	0.759	0.785	0.827
		Yes	0.837	0.876	0.856	
	80:20 (Gain Ratio)	No	0.825	0.752	0.787	0.831
		Yes	0.835	0.887	0.86	
	80:20 (Gini Index)	No	0.825	0.769	**0.796**	0.837
		Yes	0.844	0.885	**0.864**	

Table 7. **Evaluation of predictive models: moratorium.**

Machine Learning Technique	Partitioning Ratio	Target Variable: *Moratorium*	Precision	Sensitivity	F-score	Accuracy
Decision Tree	80:20(Gini Index)	No	0.779	0.727	0.752	0.802
		Yes	0.816	0.854	0.834	
Gradient Boosted Tree	60:40	No	0.891	0.711	0.791	0.844
		Yes	0.821	0.938	0.876	
Random Forest	80:20(Gini Index)	No	0.825	0.769	0.796	0.837
		Yes	0.844	0.885	0.864	
Naïve Bayes	60:40	No	0.64	0.453	0.53	0.667
		Yes	0.679	0.819	0.742	

data set is small, k-fold cross validation produces a superior model; but, when the data set is large, it produces no change. A current study backs up this conclusion (Marcot & Hanea, 2021)

A similar process was carried out to develop machine learning models for utility discounts and EPF and PRS subsidies. The results show that for both subsidies, Gradient Boosted Tree was identified as the best machine learning technique. Table 8 and Table 9 show that this technique can predict utility discount with 86% sensitivity, 82.1% precision and 84% F-score, as well as EPF and PRS with 83.6% sensitivity, 81.2% precision and 82.4% F-score, respectively.

Discussion

1. How can households that favour moratorium subsidies be identified?

To answer this research question, four classification models have been built using decision tree, gradient boosted tree, random forest and naïve bayes machine learning techniques. The optimal model from each of these techniques are determined by tuning their parameters and the details of parameter values are explained in the previous section. Finally, the best model from each of the four machine learning algorithms is compared, and the best model is chosen using the F-score performance evaluation measure. When the data division ratio is 60 percent training data and 40 percent testing data, Gradient Boosted Tree was shown to be the best machine learning model for predicting moratorium subsidies preferred households, with F-score =0.876 and sensitivity = 0.938. Hence this model is recommended for the deployment phase.

Although the gradient boosting tree can more accurately identify households that favour moratorium subsidies, it is difficult to interpret. It's because the relationship between each predictor and the target is modelled using a curve, making it difficult to explain how each predictor relates to the target. Machine learning techniques are always a trade-off between prediction accuracy and interpretability. In general, a method's interpretability reduces as its accuracy improves. (Hastie *et al.*, 2021). Therefore, the decision tree model is utilised to create the ruleset in order to determine the general profile of families who favour a moratorium subsidy. Rule support refers to the number of respondents to whom this condition applies. Rule confidence indicates the probability of having a moratorium as the preferred subsidy. Table 10 shows the basic characteristics of families that choose moratorium subsidies with a rule support of 400 and above.

Table 8. Evaluation of predictive models: utilities discount.

Machine Learning Technique	Partitioning Ratio	Target Variable: *Utilities Discount*	Precision	Sensitivity	F-score	Accuracy
Decision Tree	80:20 (Gini Index)	No	0.849	0.835	0.842	0.83
		Yes	0.809	0.824	0.816	
Gradient Boosted Tree	80:20	No	0.876	0.841	0.859	0.85
		Yes	0.821	0.86	0.84	
Random Forest	80:20 (Information Gain Ratio)	No	0.848	0.864	0.856	0.843
		Yes	0.836	0.817	0.827	
Naïve Bayes	60:40	No	0.643	0.623	0.633	0.609
		Yes	0.571	0.592	0.581	

Table 9. Evaluation of predictive models: EPF and PRS withdrawals.

Machine Learning Technique	Partitioning Ratio	Target Variable: *EPF & PRS*	Precision	Sensitivity	F-score	Accuracy
Decision Tree	80:20 (Gini Index)	No	0.882	0.854	0.868	0.844
		Yes	0.791	0.828	0.809	
Gradient Boosted Tree	60:40	No	0.889	0.871	0.88	0.857
		Yes	0.812	0.836	0.824	
Random Forest	60:40 (Information Gain Ratio)	No	0.87	0.892	0.881	0.855
		Yes	0.831	0.799	0.815	
Naïve Bayes	60:40	No	0.699	0.781	0.738	0.666
		Yes	0.599	0.494	0.541	

Table 10. General profiles for moratorium subsidy preference.

Condition	Rule Support	Rule Confidence
$Q12: Cash Assistance$ IN (“Yes”) AND $Q12: Utilities Discount$ IN (“No”) AND $Q12: EPF & PRS$ IN (“No”) AND $Q12: Wage Sub$ IN (“Yes”) AND $Q12: CGS$ IN (“No”) AND $Q10: Race$ IN (“Malay”, “Native Sabah/Sarawak”, “Others”) AND $Q7: Age Group$ IN (“35-44 “, “25-34 “, “45-54 “, “55-64 “)	902	95.68%
$Q22: Outbreak lifestyle changes?$ IN (“Yes”) AND $Q4: Gender$ IN (“Male”) AND $Q34: Internet access lvl$ IN (“NA”, “Fast”) AND $Q7: Age Group$ IN (“35-44 “) AND $Q5: Area$ IN (“City”) AND $Q7: Age Group$ IN (“35-44 “, “25-34 “, “45-54 “) AND $Q8: Marital Status$ IN (“Married”) AND $Q12: Wage Sub$ IN (“No”) AND $Q12: CGS$ IN (“No”) AND $Q10: Race$ IN (“Malay”, “Native Sabah/Sarawak”, “Others”) AND $Q7: Age Group$ IN (“35-44 “, “25-34 “, “45-54 “, “55-64 “)	792	88.51%
$Q23: Readiness of lifestyle changes$ IN (“Yes”) AND $Q12: EPF & PRS$ IN (“Yes”) AND $Q12: Utilities Discount$ IN (“No”) AND $Q22: Outbreak lifestyle changes?$ IN (“Yes”) AND $Q2: State$ IN (“Selangor”, “Perak”, “Sarawak”, “Kedah”, “KL”, “Johor”, “Perlis”, “Putrajaya”, “Pahang”, “Melaka”, “Pulau Pinang”, “Terengganu”, “Negeri Sembilan”, “Sabah”, “Labuan”) AND $Q4: Gender$ IN (“Male”) AND $Q18: Future ESP?$ IN (“Yes”) AND $Q12: Cash Assistance$ IN (“Yes”) AND $Q5: Area$ IN (“Rural”) AND $Q7: Age Group$ IN (“35-44 “, “25-34 “, “45-54 “) AND $Q8: Marital Status$ IN (“Married”) AND $Q12: Wage Sub$ IN (“No”) AND $Q12: CGS$ IN (“No”) AND $Q10: Race$ IN (“Malay”, “Native Sabah/Sarawak”, “Others”) AND $Q7: Age Group$ IN (“35-44 “, “25-34 “, “45-54 “, “55-64 “)	466	93.13%
$Q12: EPF & PRS$ IN (“Yes”) AND $Q12: Utilities Discount$ IN (“No”) AND $Q12: Cash Assistance$ IN (“Yes”) AND $Q34: Internet access lvl$ IN (“NA”, “Slow”, “Average”, “No Internet Access”) AND $Q22: Outbreak lifestyle changes?$ IN (“Yes”) AND $Q4: Gender$ IN (“Female”) AND $Q2: State$ IN (“Selangor”, “Kedah”, “KL”, “Johor”, “Perlis”, “Putrajaya”, “Pulau Pinang”, “Negeri Sembilan”, “Sabah”) AND $Q7: Age Group$ IN (“25-34 “, “45-54 “, “15-24 “, “55-64 “, “65 and above”) AND $Q5: Area$ IN (“City”) AND $Q7: Age Group$ IN (“35-44 “, “25-34 “, “45-54 “) AND $Q8: Marital Status$ IN (“Married”) AND $Q12: Wage Sub$ IN (“No”) AND $Q12: CGS$ IN (“No”) AND $Q10: Race$ IN (“Malay”, “Native Sabah/Sarawak”, “Others”) AND $Q7: Age Group$ IN (“35-44 “, “25-34 “, “45-54 “, “55-64 “)	447	90.16%

F1000Research 2021, 10:932 Last updated: 03 DEC 2021

The first rule shows that households who prefer to have a cash allowance and their race is either Malay or Native Sabah/ Sarawak or others, while those aged between 25 to 64 prefer moratorium. Table 11 explains the first rule indicating the general profile of households who prefer moratorium subsidies.

2. How to find out which households seek utility discount subsidies?

The four machine learning techniques described in the moratorium subsidies were applied to develop the classification model in order to identify the households who want utility discount subsidies. All of the procedures outlined in the preceding section were followed in order to find the optimal machine learning model. The gradient boosting tree outperforms the other three techniques, with a data partitioning ratio of 80:20, an F-score of 0.84, and a sensitivity of 0.86. Although the gradient boosting tree can more accurately identify households who seek utilities discount subsidies, it is difficult to interpret this model. As a result, decision tree rules were developed in order to comprehend the overall profile of households who seek utility discount subsidies. One such rule is presented in Table 12.

3. How can households who desire EPF and PRS withdrawal subsidies be identified?

To identify the households who want EPF and PRS withdrawal subsidies., similar to the moratorium and utilities discount subsidies classification models, decision tree, gradient boosted tree, random forest and naïve bayes techniques were used to develop the machine learning model. With a data partitioning ratio of 60:40 and F-scores of 0.824 and sensitivity of 0.836, the gradient boosting tree was found to be the best model compared to the others. Decision tree rules were developed to explain the general features of households who prefer EPF and PRS withdrawal subsidies, and one of the rules is displayed in Table 13.

The results imply that households that prefer moratorium subsidies did not favour other financial aids except cash assistance. By contrast, households that prefer for utility discounts,

Table 11. General profile of moratorium subsidy preference - rule support: 902 records, rule confidence: 95.7%.

Subsidies	Race	Age Group
• Cash Assistance (Yes) • Utilities Discount (No) • EOF & PRS (No) • Wage Subsidies (No) • CGS (No)	• Malay • Native Sabah / Sarawak • Others	• 35-44 • 25-34 • 45-54 • 55-64

Table 12. General profile of utilities discount subsidy preference - rule support: 1128 records, rule confidence: 90.1%.

Readiness of Lifestyle Changes	Marital Status	States	Subsidies
Yes	• Married • Divorced / Separated	• KL • Putrajaya • Melaka • Pulau Penang • Sarawak • Johor	• CGS (No) • Cash Assistance (Yes) • Wage Subsidies (No) • Moratorium (Yes) • EPF & PRS (No)

Table 13. General profile of EPF & PRS subsidy preference - rule support: 1427 records, rule confidence: 90.5%.

ESP in future	Marital status	Priority	Race	States	Subsides
Yes	• Single / Non-Married • Married • Widow / Widower • Divorced / Separated	Employment	• Malay • Native Sabah / Sarawak • Others • Foreigner	• KL • Selangor • Johor • Labuan • Sarawak • Negeri Sembilan	• CGS (No) • Cash Assistance (Yes) • Wage Subsidies (No) • Moratorium (Yes) • Utilities Discount (No)

EPF and PRS withdrawals also chose moratorium subsidies and cash assistance. All households preferred cash assistance, which had the highest score among financial aids, followed by moratorium subsidies. Utility discounts, EPF and PRS withdrawals can be implemented according to the household income group preferences.

Wage subsidy and credit guarantee scheme were the least preferred financial assistance. First, the Prihatin wage subsidy is only for eligible Social Security Organisation (SOCSO) subscribers. Hawkers, small businesses and their employees might not subscribe to SOCSO and thus, ineligible to apply for financial aid. Second, the credit guarantee scheme was not preferred due to economic uncertainty from COVID-19. Economic uncertainty adversely affects household income, resulting in their inability to repay the loan instalments.

Limitations of the study

The following are some of the limitations of this study's findings: The data used are survey responses and cannot be considered to represent the views of all Malaysians. According to DOSM, it should not be used to analyse the impact of COVID-19 in Malaysia and should not be considered official statistics. It can, however, be utilised to assist in the reflection process ("Department of Statistics Malaysia Official Portal," 2020; Malaysia, 2020). Another limitation is data partitioning method. This study portioned the data into training and test data sets. However, to improve the model, the dataset could be divided into training, validation and test data. This research used information based learning (a decision tree algorithm family) as the best model for explaining decision logic. However, it is possible to test with other classification methods, and they might have better accuracy.

Conclusions

This study used data analytics and machine learning approaches to derive insights from the "Effects of COVID-19 on the Economy and Individual - Round 2" survey dataset. The CRISP-DM approach was applied to develop prediction models for households' preferred subsidies, such as moratoriums, utility discounts and EPF and PRS using four machine learning algorithms, namely, Decision Tree, Random Forest, Naïve Bayes and Gradient Boosted Tree. For all three subsidies, the best predictive model was obtained by Gradient Boosted Tree. The findings can be used to design customised ESPs that effectively manage the economic burden of low-income households.

Data availability

Data used in this study were obtained from a survey dataset "Effects of COVID-19 on the Economy and Individual - Round 2," available from the Department of Statistics, Malaysia (DOSM). A report published by the DOSM based on the survey can be viewed on the DOSM website. Access to this data requires application, as stated on the DOSM website. A guide for how to apply for dataset access is available on the Data Request page or requests for more information can be emailed to data@dosm.gov.my.

Acknowledgements

The authors thank the Department of Statistics Malaysia for allowing us to use "Effects of COVID-19 on the Economy and Individual - Round 2" survey data to carry out this study. The authors also acknowledge Multimedia University for supporting this research.

References

Department of Statistics Malaysia Official Portal: 2020; Retrieved June 27, 2021.
Reference Source

Flanders S, Nungsari M, Chuah HY: **The COVID-19 Hardship Survey: An Evaluation of the Prihatin Rakyat Economic Stimulus Package**. 2020; 1–44.
Reference Source

Hastie T, Tibshirani R, James G, *et al.*: **An Introduction to Statistical Learning (2nd Edition).** *Springer Texts.* 2021; **102**: 618.
Reference Source

Kelleher JD, Namee BM, D'Arcy A, *et al.*: **Fundamentals of Machine Learning for Predictive Data Anayltics: Algorithms, Worked Examples, and Case Studies**. (Ist). The MIT Press, 2015.
Reference Source

Malaysia D of S: **Technical Note _Special Survey Covid19_Individuals**. 2020.

Marcot BG, Hanea AM: What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Comput Stat.* 2021; **36**(3): 2009–2031.
Publisher Full Text

Moscato V, Picariello A, Sperlí G: **A benchmark of machine learning approaches for credit score prediction.** *Expert Syst Appl.* 2021; **165**: 113986.
Publisher Full Text

Mostafa MM, El-Masry AA, Chen Y, *et al.*: **Impact of Classification Algorithms on Census Dataset.** *International Journal of Recent Technology and Engineering.* 2021; **97**(2019): 526–534.

Sangavi N, Jeevitha R, Kathirvel P, *et al.*: **Impact of Classification Algorithms on Census Dataset.** *International Journal of Recent Technology and Engineering.* 2020; **8**(5): 2666–2670.
Publisher Full Text

Schröer C, Kruse F, Gómez JM: **A Systematic Literature Review on Applying CRISP-DM Process Model.** *Procedia Comput Sci.* 2021; **181**: 526–534.
Publisher Full Text

Shah AUM, Safri SNA, Thevadas R, *et al.*: **COVID-19 outbreak in Malaysia: Actions taken by the Malaysian government.** *Int J Infect Dis.* 2020; **97**: 108–116.
PubMed Abstract | Publisher Full Text | Free Full Text

Trivedi SK: **A study on credit scoring modeling with different feature selection and machine learning approaches.** *Technology in Society.* 2020; **63**: 101413.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ? ✓

Version 2

Reviewer Report 03 December 2021

https://doi.org/10.5256/f1000research.79204.r100079

© **2021 N.K S.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✓ **Sreeja N.K**
Department of Applied Mathematics and Computational Sciences, PSG College of Technology, Coimbatore, Tamil Nadu, India

The authors have addressed the issues raised.
The authors have reported the average results of cross validation.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Pattern recognition, Supervised Learning, Swarm Intelligence

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 25 October 2021

https://doi.org/10.5256/f1000research.76592.r97108

© **2021 N.K S.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Sreeja N.K**
Department of Applied Mathematics and Computational Sciences, PSG College of Technology, Coimbatore, Tamil Nadu, India

- The authors have compared few classifiers on Effects of COVID-19 on the Economy and

Individual - Round 2 data set to predict individual household preferences from ESP.

- I would recommend the authors to use k-fold cross validation method to evaluate the performance of a classifier instead of a random 60-40 or 80-20 split. This would guarantee that all k subsets are used as a test set once.

- The authors can also perform some statistical tests to validate the results obtained from experiments.

- In Page 4, (section Evaluation), TPR is the True positive rate and not as mentioned by the authors.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
No

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Supervised Learning, Swarm Intelligence

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 12 Nov 2021

Rathimala Kannan null, Multimedia University, Cyberjaya, Malaysia

- I would recommend the authors to use k-fold cross validation method to evaluate the performance of a classifier instead of a random 60-40 or 80-20 split. This would guarantee that all k subsets are used as a test set once.

Thank you for your comments. As we are aimed to tune the parameters to get the optimal model from each machine learning techniques, separate tables were provided

with the various parameter values. Also the data size was 40k.
We conducted K-fold cross validation method to evaluate the performance of a classifier. When the data is partitioned with k= 5, and K=10, the findings reveal little difference in F-score values, and the Gradient boosting tree performs best. When the data set is small, k-fold cross validation produces a superior model; but, when the data set is large, it produces no change. A current study backs up this conclusion (Marcot & Hanea, 2021).

Marcot, B. G., & Hanea, A. M. (2021). What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Computational Statistics*, *36*(3), 2009–2031. https://doi.org/10.1007/s00180-020-00999-9

Machine Learning Technique	**Partitioning Ratio**	**Target Variable: *Moratorium***	**Precision**	**Sensitivity**	**F-score**	**Accuracy**
Decision Tree	80:20	(Gini Index)	No	0.779	0.727	0.752	0.802
			Yes	0.816	0.854	0.834	
	K=5-fold cross validation	(Gini Index)	No	0.770	0.702	0.735	0.789
			Yes	0.801	0.851	0.826	

K=10-fold cross validation
(Gini Index)
No
0.772
0.697
0.733
0.789

Yes
0.799
0.854
0.826

Gradient Boosted Tree
60:40
No
0.891
0.711
0.791
0.844

Yes
0.821
0.938
0.876

K=5-fold cross validation
No
0.889
0.705
0.786
0.841

Yes
0.818
0.938
0.874

K=10-fold cross validation
No
0.889
0.705
0.787
0.841

Yes

0.818
0.938
0.874

Random Forest
80:20
(Gini Index)
No
0.825
0.769
0.796
0.837

Yes
0.844
0.885
0.864

K=5-fold cross validation
(Gini Index)
No
0.821
0.761
0.790
0.832

Yes
0.839
0.882
0.860

K=10-fold cross validation
(Gini Index)
No
0.826
0.757
0.790
0.833

Yes
0.837
0.887
0.862

Naïve Bayes
60:40
No

0.64
0.453
0.53
0.667

Yes
0.679
0.819
0.742

K=5-fold cross validation
No
0.634
0.446
0.524
0.664

Yes
0.676
0.818
0.740

K=10-fold cross validation
No
0.634
0.446
0.524
0.664

Yes
0.676
0.818
0.740

- In Page 4, (section Evaluation), TPR is the True positive rate and not as mentioned by the authors.

Thank you for spotting this typo error. We have corrected it.

Competing Interests: No

Reviewer Report 15 October 2021

https://doi.org/10.5256/f1000research.76592.r95431

© 2021 Ismail S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Shahrinaz Ismail
Albukhary International University, Alor Setar, Kedah, Malaysia

Since this is a full research paper, it is recommended that a Literature Review section be included. It is believed that there are some previous research that have performed the 4 ML techniques covered by this research, hence the need to also mention those and how they differ in results as compared to the results found in here.

Some additional elaboration is required for technical parts of this paper, especially in clarifying the relation between the formula and the techniques used. Do all the formulas need to be performed for each of the 4 techniques?

If there is a step taken before the evaluation that uses the formulas, please include them at least as summary. If this part is covered in the published paper (part 1), perhaps that should be mentioned. If this paper only covers the result on which technique is the best, then perhaps the title should reflect that. At the moment, the title does not reflect the research questions.

Research questions are usually asked with the word "how" instead of "can". It would be more sound to rephrase them to "how".

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Data analytics with knowledge management

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 12 Nov 2021

Rathimala Kannan null, Multimedia University, Cyberjaya, Malaysia

Since this is a full research paper, it is recommended that a Literature Review section be included. It is believed that there are some previous research that have performed the 4 ML techniques covered by this research, hence the need to also mention those and how they differ in results as compared to the results found in here.
We have added brief literature review in the modelling section. Due to the limitation of total words to be within 2500 words, a separate section is not added.

Some additional elaboration is required for technical parts of this paper, especially in clarifying the relation between the formula and the techniques used. Do all the formulas need to be performed for each of the 4 techniques?
We employed standard performance indicators to evaluate the classification models.

If there is a step taken before the evaluation that uses the formulas, please include them at least as summary. If this part is covered in the published paper (part 1), perhaps that should be mentioned. If this paper only covers the result on which technique is the best, then perhaps the title should reflect that. At the moment, the title does not reflect the research questions.
We have revised the discussion section which summarizes the process and the results.

Research questions are usually asked with the word "how" instead of "can". It would be more sound to rephrase them to "how".
The research questions are rephrased to start with "how".

Competing Interests: No

Reviewer Report 06 October 2021

https://doi.org/10.5256/f1000research.76592.r95430

© 2021 Pramana S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Setia Pramana
[1] Computational Statistics Department, Politeknik, Statistika STIS, Jakarta, Indonesia
[2] BPS Statistics Indonesia, Jakarta, Indonesia

1. Research questions and the discussion and conclusion are not inline. The authors

mentioned three research questions, but they are not discussed throughout the manuscript. Instead, the authors focus on comparing different methods for predicting household subsidy preferences. Please address the research question on the results and discussion as well.

2. Please explain the question used in the study and why some questions are missing. Is it due to not being the focus of the study, or for other reasons?

3. Please elaborate more on the discussion of the results of the study. What would be the impact of utilizing best methods for result of the stimulus, especially for the citizen? Would it be that big a difference if we use the other classification methods?

4. Please explain the advantages and disadvantages of the methods being compared in the study.

5. The result shows that the decision tree, gradient boost, and random forest, does not differ much, how will it affect the stimulus package?

6. If the data changes would the results change? The authors may divide the data into testing, training and validation data set. Several times as discussed by this paper: Lessman *et al.* (2015[1]).

7. Please use a flowchart to explain the modelling and analysis process.

8. Why discuss on different ratio partitioning, in this case it is better use simulation dataset to see the impact of the ratio into the results.

References

1. Lessmann S, Baesens B, Seow H, Thomas L: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*. 2015; **247** (1): 124-136 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine Learning, Data mining, statistics, official statistics, bioinformatics, biostatistics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 12 Nov 2021

Rathimala Kannan null, Multimedia University, Cyberjaya, Malaysia

1. Research questions and the discussion and conclusion are not inline. The authors mentioned three research questions, but they are not discussed throughout the manuscript. Instead, the authors focus on comparing different methods for predicting household subsidy preferences. Please address the research question on the results and discussion as well.

Though we have answered all three research questions in the discussion section, we noticed it is not presented well. Thank you for your comment. Now we have answered each question separately in the discussion section.

2. Please explain the question used in the study and why some questions are missing. Is it due to not being the focus of the study, or for other reasons?

As mentioned in the previous answer, now we have revised the discussion section which clearly addresses all three research questions.

3. Please elaborate more on the discussion of the results of the study. What would be the impact of utilizing best methods for result of the stimulus, especially for the citizen? Would it be that big a difference if we use the other classification methods?

The aim of this study is to identify the households that prefer to have a specific subsidy which is expected to lessen their financial burden during this pandemic. Therefore, we focus on the classification model which can predict more accurately the household's choice.
It is possible to test with other classification methods, and they might have better accuracy. However, we focus also the model interpretability, thus we use mostly on decision tree algorithm family plus naïve bayes algorithm. Machine learning models are always a trade-off between accuracy and interpretability (Hastie, Tibshirani, James, & Witten, 2021).

Hastie, T., Tibshirani, R., James, G., & Witten, D. (2021). An Introduction to Statistical

Learning (2nd Edition). *Springer Texts*, *102*, **618. (chapter 2.1.3 page 24)**

4. Please explain the advantages and disadvantages of the methods being compared in the study.

Different approaches, such as information-based learning, similarity-based learning, error-based learning, and probability-based learning, can be used to create classifiers. This research used information based learning (a decision tree algorithm family) as the best model for explaining decision logic .(Kelleher, Namee, & D'Arcy, 2015) Kelleher, J. D., Namee, B. Mac, & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Anayltics: Algorithms, Worked Examples, and Case Studies* (Ist). The MIT Press.

5. The result shows that the decision tree, gradient boost, and random forest, does not differ much, how will it affect the stimulus package?

The aim here is to predict household's preference as much possible correctly. Thus we chose the model that performs the best compared to the rest.

6. If the data changes would the results change? The authors may divide the data into testing, training and validation data set. Several times as discussed by this paper: Lessman *et al.* (2015[1]). –

Yes. We can divide the dataset in to 3 sets. We conducted K-fold cross validation method to evaluate the performance of a classifier and presented the results in the following table. When the data is partitioned with k= 5, and K=10, the findings reveal little difference in classifier performance, and the Gradient boosting tree performs best. When the data set is small, k-fold cross validation produces a superior model; but, when the data set is large, it produces no change. A current study backs up this conclusion (Marcot & Hanea, 2021).

Marcot, B. G., & Hanea, A. M. (2021). What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Computational Statistics*, *36*(3), 2009–2031. https://doi.org/10.1007/s00180-020-00999-9

Machine Learning Technique
Partitioning Ratio
Target Variable: *Moratorium*
Precision
Sensitivity
F-score
Accuracy

Decision Tree
80:20
(Gini Index)

No
0.779
0.727
0.752
0.802

Yes
0.816
0.854
0.834

K=5-fold cross validation
(Gini Index)
No
0.770
0.702
0.735
0.789

Yes
0.801
0.851
0.826

K=10-fold cross validation
(Gini Index)
No
0.772
0.697
0.733
0.789

Yes
0.799
0.854
0.826

Gradient Boosted Tree
60:40
No
0.891
0.711
0.791
0.844

Yes
0.821

0.938
0.876

K=5-fold cross validation
No
0.889
0.705
0.786
0.841

Yes
0.818
0.938
0.874

K=10-fold cross validation
No
0.889
0.705
0.787
0.841

Yes
0.818
0.938
0.874

Random Forest
80:20
(Gini Index)
No
0.825
0.769
0.796
0.837

Yes
0.844
0.885
0.864

K=5-fold cross validation
(Gini Index)
No
0.821
0.761
0.790

0.832

Yes
0.839
0.882
0.860

K=10-fold cross validation
(Gini Index)
No
0.826
0.757
0.790
0.833

Yes
0.837
0.887
0.862

Naïve Bayes
60:40
No
0.64
0.453
0.53
0.667

Yes
0.679
0.819
0.742

K=5-fold cross validation
No
0.634
0.446
0.524
0.664

Yes
0.676
0.818
0.740

K=10-fold cross validation
No

0.634
0.446
0.524
0.664

Yes
0.676
0.818
0.740

7. Please use a flowchart to explain the modelling and analysis process.
A flowchart has been provided. Thank you for your comment.

8. Why discuss on different ratio partitioning, in this case it is better use simulation dataset to see the impact of the ratio into the results.

Data partitioning ratio is considered as one of the parameters and the aim was to identify the best ratio that can provide better classification model. We also used k-fold cross validation method as explained earlier.

Competing Interests: No competing interest

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research