Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol 5, No.1, March 2022, pp. 54 – 64
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Identifying Characteristics of Households Recipient of the Government's Social Protection Program

[1]Nofrida Elly Zendrato, [2]Bagus Sartono, [3]Utami Dyah Syafitri
[1,2,3]Department of Statistics, IPB University
Email: [1]nofrida.nofrida@apps.ipb.ac.id, [2]bagusco@apps.ipb.ac.id, [3]utamids@apps.ipb.ac.id

Article Info

Article history:
Received Jul 21th, 2022
Revised Aug 04th, 2022
Accepted Aug 18th, 2022

Keyword:
Permutation feature importance
Random Forest
SHAP
SMOTE
Social Protection

ABSTRACT (10 PT)

The percentage of poor people in Indonesia increased in March 2021 compared to March 2020, as is the condition in Banten Province in the last three years. One way the government can overcome poverty, equality, and other transformations in the lower middle class are by providing social protection programs. This study seeks to examine the characteristics of households receiving social protection programs. The data used is the National Socio-Economic Survey in March 2021, Banten Province. The model's method is a random forest, followed by the permutation feature importance and Shapley additive explanation method to obtain important variables. Important variables were selected based on consistent top rankings in both methods. Before forming the random forest model, the data imbalances in the response variables were handled using the SMOTE technique. Evaluation of the classification model obtained an AUC value of 0,718. Shapley additive explanation is more consistent than the importance of permutation features. Six important variables, namely capita expenditure, education of the head of the household, age of the head, source of drinking water, floor area, and the number of household members.

Corresponding Author:
Nofrida Elly Zendrato
Department of Statistics,
IPB University,
Raya Dramaga Road, Babakan, Dramaga 16680, Bogor, Indonesia.
Email: nofrida.nofrida@apps.ipb.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v5i1.18579

1. INTRODUCTION

According to Statistics Indonesia, poor people have an average monthly per capita expenditure below the poverty line. The percentage of poor people in March 2021 increased by 1,12 million compared to March 2020, which was 10,14 percent but decreased by 0,05 percent against September 2020. Several factors of poverty/food insecurity are the Gini ratio, open unemployment rate, recipients of National Health Insurance Contribution Assistance Programs [1], program recipients, Raskin recipients, education level [2], cooking fuel, widest floor type, education of household, defecation facility, wall types, water sources, and regional status [3]. Some of these variables are used in this study, and research [2] [3] became the main reference in this study.

Reducing poverty is one of the government's targets by providing social protection programs that include social security and social assistance programs. Some programs include the Prosperous Family Card, Hope Family Program, Smart Indonesia Card, Smart Indonesia Card, food aid, pre-employment card, and other local government assistance. Banten Province was one of the five provinces with the highest number of National Health Insurance Contribution Assistance Programs in Indonesia [1].

Target recipients of social protection programs are closely related to poverty and food insecurity. The distribution of these programs is expected to help fulfill the minimum basic needs of a person, family, and lower-middle-class constructed community. Therefore, a deeper study is needed regarding the characteristics of households receiving social protection programs so that they become input for policymakers.

Modeling is carried out to classify households receiving social protection programs using one of the techniques in machine learning. Machine learning is the part of artificial intelligence that is more popular and

has experienced many developments recently. Classification techniques are part of machine learning. One of the popular classification techniques is the random forest, namely the development of a decision tree and the application of the bagging method (bootstrap and aggregating) in forming a classification tree. The random forest forms a classification tree that is independent of other trees.

Some researchers state that the performance of the random forest method is superior to other machine learning methods. Random forest accuracy is better than Partial Least Discriminant Analysis (PLS-DA), Support Machine Learning (SVM), and Voting Feature Interval 5 (VFI 5) in predicting the efficacy of herbal medicine [4]. Random forest accuracy is better than Adaboost in predicting UKT delay [5]. The accuracy of random forest is better than the Classification and Regression Tree (CART) in the classification of success in continuing education at the high school level in Banten Province [6]. The ensemble method (random forest and AdaBoost) is better than the single classifier method (decision tree and K-nearest neighbors) [7]. It was concluded that the random forest method was used to model the status of households receiving social protection programs.

Unbalanced data results in prediction errors, so data imbalances are needed to be handled. Synthetic Minority Oversampling Technique (SMOTE) is a technique for handling unbalanced data. The basic idea is to generate new synthetic data from minority classes using the k-nearest neighbor's approach to obtain a class equivalent to the majority [8].

The interpretation technique in the popular classification model used recently is a technique to explain the predictions of any classifier model that can be interpreted and trusted. The interpretation technique is permutation feature importance and Shapley Additive Explanations (SHAP). Both methods obtain important variables based on their contribution to the response variable [9]. The consistent important variables at the top rank were further analyzed to see the relationship between the categories of important variables and the response variables by looking at the biplot.

2. RESEARCH METHOD

2.1 Literature View

2.1.1 Random Forest

Random forest is the development of a decision tree, where each tree is trained using individual examples. This technique also applied the ensemble and bagging methods (bootstrap and aggregating) in forming the classification tree so that the trees started to tend not to be similar to other trees. This will have an impact on the accuracy of the resulting better predictions. Random forest is relatively robust against outliers and noise, has a low bias, and can avoid overfitting or underfitting. The application of bootstrap sampling in building a prediction tree by combining each decision tree's results based on the most votes [10]. The random forest workflow is shown in Figure 1.

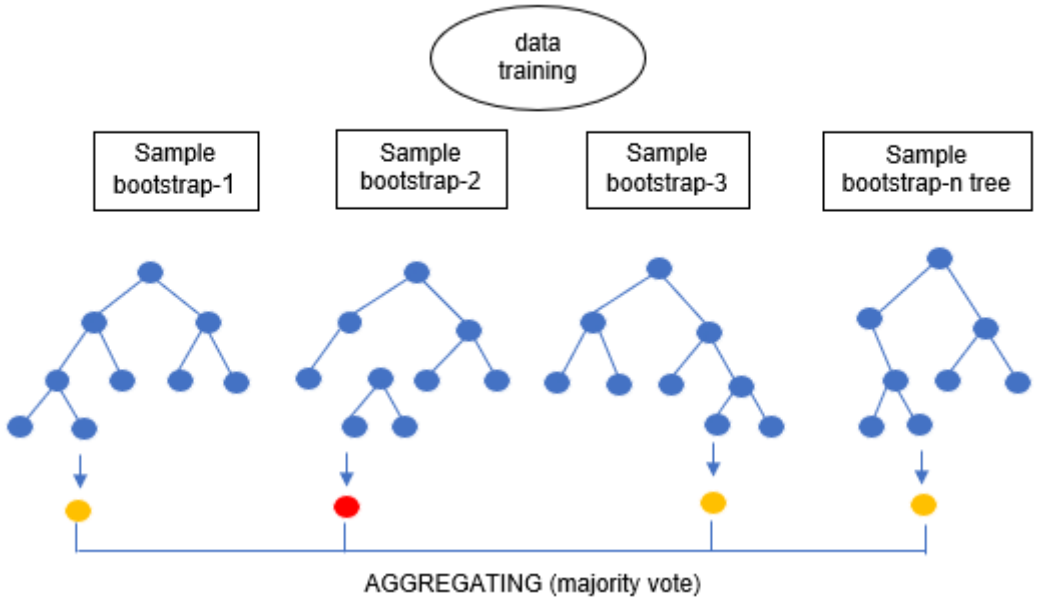

Figure 1. Basic Random Forest Algorithm

Unbalanced data can cause prediction errors, so we must handle it. SMOTE is a technique for handling data imbalances with the basic idea of generating new synthetic data from the minority class with the closest neighbor approach to obtain a class equivalent to the majority class.

The model's goodness measures are accuracy, sensitivity, and specificity. Accuracy Accuracy is used to see how accurate the model predicts. Sensitivity is used to see how accurately the model classifies the positive class (program recipient) into the positive class. Specificity is used to see how accurately the model classifies the negative class (not program recipient) into the negative class category. The ROC (receiver operator characteristics) curve can provide more information about summarizing predictive performance [11]. AUC (area under the curve) of ROC describes the performance of a classifier with values ranging from 0 to 1. On the ROC curve, the x-axis is the false positive rate (FPR), and the y-axis is the true positive rate (TPR) or sensitivity. The greater the AUC value, it can be said that the classifier model used is stronger. Accuracy is shown in equation (1), sensitivity (TPR) is shown in equation (2), and specificity (true negative rate) is shown in equation (3), and more details can be seen in Table 1.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN)}} \text{ x } 100\% \quad (1)$$

$$\text{Sensitivity (TPR)} = \frac{\text{TP}}{\text{(TP+FN)}} \quad (2)$$

$$\text{Specificity (TNR)} = \frac{\text{TN}}{\text{(TN+FP)}} \quad (3)$$

$$\text{FPR} = \frac{\text{FP}}{\text{(TN+FP)}} \quad (4)$$

Table 1. Confusion matrix

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	True Positive	False Positive
	Negative (0)	False Negative	True Negative

2.1.2 Permutation Feature Importance

Obtaining a classification model is not enough; there needs to be an interpretation technique that can explain the model. Permutation feature importance is a global interpretation technique with the basic idea of calculating the increase in the model's prediction error after changing the order of features. The algorithm used based on Fisherm Rudin and Dominici (2018):

1. Estimating the error of the original model

$$e^{orig} = L(y, f(X)) \quad (5)$$

2. For each j-feature = 1, …, p
   a. Generate the feature matrix X^{perm} with a permutation of the j-features in the X data. Breaks the relationship between the j-features and y of the actual result.
   b. Estimating the error (error) based on the predicted data permutation

$$e^{perm} = L(Y, f(X^{perm})) \quad (6)$$

   c. Calculating the importance of permutation features

$$FI^{j} = \frac{e^{perm}}{e^{orig}} \quad (7)$$

$$FI^{j} = e^{perm} - e^{orig} \quad (8)$$

3. Sort features by lowering FI

   Permutation feature importance can be applied to any machine learning model. This technique provides a very dense global insight into the model's behavior. However, the results of this technique can be biased if the variables are strongly correlated, so caution should be exercised in interpreting them [9]. Permutation feature importance is used to determine factors that have the potential to contribute to brain health in lonely individuals [22].

2.1.3. Shapley Additive Explanations (SHAP)

The SHAP technique was introduced by Lundberg [12] with the basic idea based on Shapley's value game theory. The main goal is to estimate the prediction locally by calculating the contribution of each feature to the prediction. The equation for Shapley's value in SHAP is:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_0 \, z'_j \quad (9)$$

$g(z')$ is explanation model, $z' \in \{0,1\}^M$ is coalition vector (simplified features) that 1 if feature value is present and 0 if feature value is absent, M is the maximum of coalition size, and $\phi_0$is base value from classification model.

The SHAP method is an interpretation technique that provides a complete and reasonable explanation. We can see how the model behaves globally and in individual/local behavior [9]. SHAP is used to determine the prediction of company acquisition [18]. SHAP is used to obtain important variables in routine and non-routine industrial classifications in conducting R&D [23].

2.2 Method

2.2.1 Data

The data used in this study is from Statistics Indonesia, which is the result of the National Socio-Economic Survey (Susenas) of Banten Province in March 2021. The data used consists of 27.418 individuals and 7.236 household data. The unit of observation used was the household. The variables used in this study are shown in Table 2.

Table 2. Variables

Variable	Indicator	Description
Y		Household status of social protection program acceptance
X1	Demographics	Area Status
X2		Marital status of the head of the household
X3		Gender of the head of the household
X4		Age of head of household
X5		Number of household members
X6	Demographics	Number of families living
X7	Education	Head of household education level
X8	Employment	Working status of the head of the household
X9	Housing area	Residential building ownership status
X10		The floor area of a residential building
X11		Building materials over the widest house
X12		Widest floor main building material
X13		Type of house roof
X14	Housing support	The main source of water for drinking
X15		Defecation facility
X16		The main source of home lighting
X17		The main type of cooking fuel
X18	Ownership of goods	AC ownership
X19		Car ownership
X20		Gold/jewelry ownership
X21		Land ownership
X22	Consumption	Total expenditure
X23		Per capita expenditure

The category of household status receiving social protection programs is 1, namely Yes (as recipients of social protection programs) on the condition that households receive at least one social protection program, and 0, which is No (not as recipients of social protection programs). The types of social protection programs covered in this study are Prosperous Family Card, Hope Family Program, social assistance for the elderly, assistance for disabilities, food assistance (Non-Cash Food Assistance/Sembako Program), routine, and non-routine assistance/social assistance/subsidies from local governments.

2.2.2 Data Analysis

First, pre-processing data. These steps taken are aggregating individual data at a household level, checking missing values, categorizing the social protection program acceptance status, namely 1 = Yes (as a recipient of a social protection program, namely receiving at least one program) and 0 = No (not as a recipient of a social protection program), data exploration to see an overview of the variables to be analyzed. At this stage, the description of the data class on the response variable is seen, and using the SMOTE technique on the training data to handle imbalanced data and splitting the data (training dataset 70% and validation dataset 30%).

The second is, the classification model. This step taken is to build a classification of recipients and non-recipients of the government's social protection program 100 times using random forest. Determination of

optimal parameters with 10-fold cross-validation, looking for hyperparameter random forest model by tuning the parameters to the parameters, namely: n estimators, max features, max depth, min samples leaf, and min samples split. Evaluate the performance model by accuracy, sensitivity, specificity, and AUC (Area Under Curve) ROC (Receiver Operating Characteristics).

Third, variable importance. The variables important selected at this stage are important variables that are both indicated/intersected as important features with a high level of importance in both methods (PFI and SHAP). These steps calculate the importance of permutation features and sort the importance of features, calculate Shapley values and the importance of features, and determine the important features of PFI and SHAP results.

Fourth, PCA Biplot. Principal component biplot analysis (PCA Biplot) or classical biplot is a descriptive statistical technique in the form of graphical representation that can simultaneously present *n* objects and *p* variables in one two-dimensional graph. The important variables resulting from both PFI and SHAP methods will be seen in the relationship between each category class on each important variable with biplot visualization.

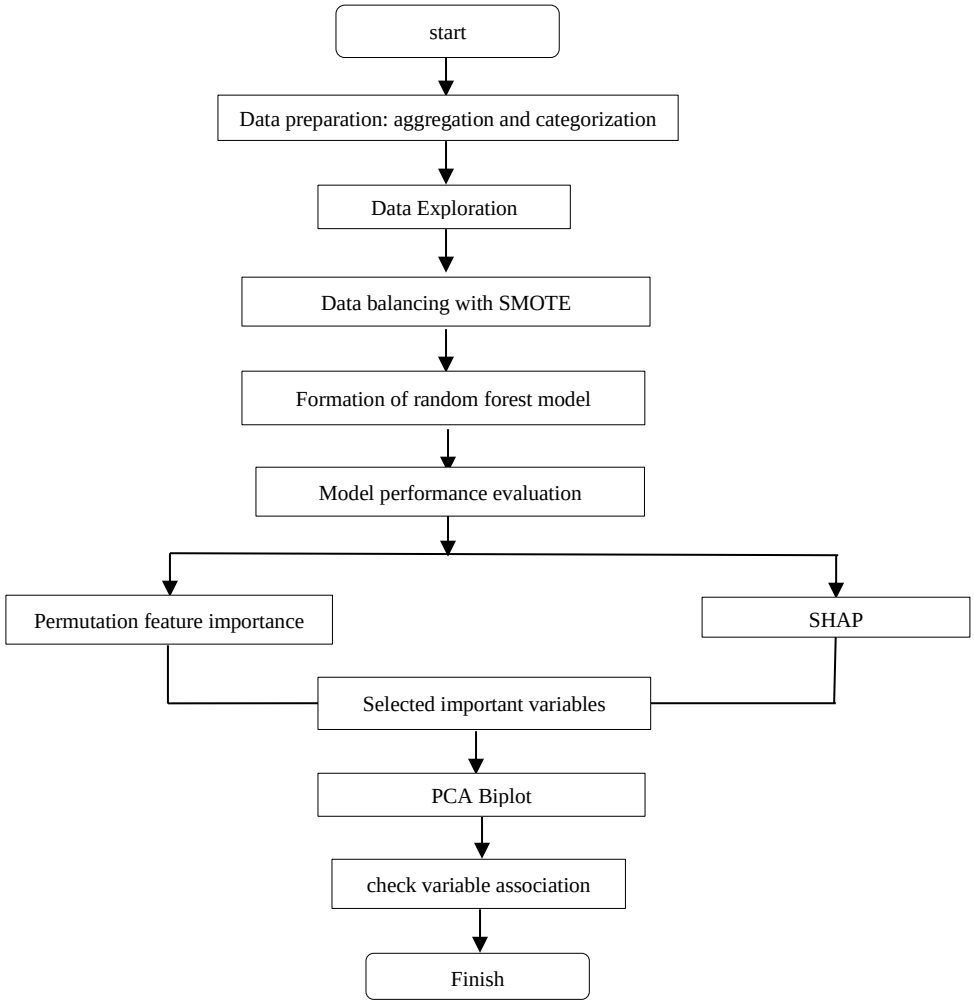

Figure 2. Flowchart

4. RESULTS AND ANALYSIS

The sample household data from National Socio-Economic Survey (Susenas) in March 2021 for Banten Province was 7.236 households, with the proportion of household status being 30% (2.149 households) as recipient and 70% (5.087 households) as not program recipient. Based on these proportions, it is shown that the data on the acceptance status of social protection programs (recipient and not recipient) tend to be unbalanced. In the classification model, both the minority class and the majority class are treated equally important, even though the minority class is often the primary concern in research [13].

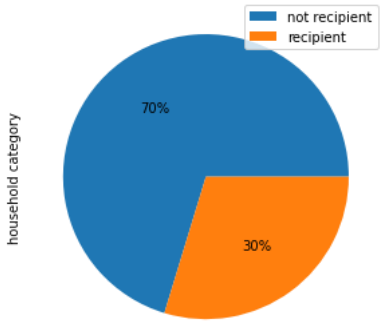

Figure 3. Household status of social protection program acceptance

3.1. Analysis of Numerical Explanatory Variables

The data distribution shows that the median values for the number of families living, floor area, age of head of household, and total expenditure tend to be higher than the average values for data on the number of families, floor area, and age of the head of household and total expenditure. Meanwhile, the data distribution on the number of household members and per capita expenditure has a median value that tends to be smaller than the average value.

Table 3. Descriptive statistics of numeric variables

Variable	Min	Mean	Median	Max
Number of families living	0	1	1.253	7
The floor area of a residential building	3	63	73.76	500
Number of household members	1	4	3,789	13
Age of head of household	16	46	47,35	97
Total expenditure	417.933	4.394.214	5.650.492	12.3719.583
Per capita expenditure	251.066	1.651.023	1.218.871	2.0619931

3.2. Analysis of Categorical Explanatory Variables

Characteristics of the majority of sample households were urban, own house, the type of roof was tile, the type of floor was marble/ceramic, the widest type of wall was the wall, had defecation facility, source of drinking water was bottled/refillable water, electricity was PLN without a meter, cooking fuel was 3 kg LPG, ownership of air conditioners, cars, gold was not owning, own land, and characteristics of household head were male, graduated from primary school/equal, work and marry.

Identification of Categorical Explanatory Variables Based on Demography

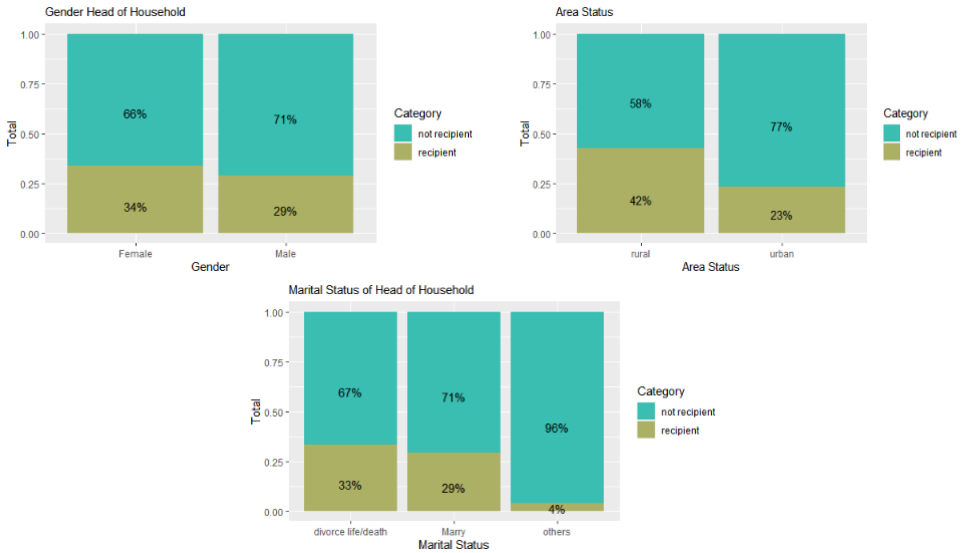

Figure 4. The Proportion of Categorical Variables Based on Demography

This stage is carried out to identify whether the variable can explain the response variable well. Variables showing differences are suggested to enter into the following modeling process. From Figure 4, it is shown that the proportion of household status shows differences in the explanatory variables of area status and

marital status of the head of the household. In contrast, the gender of the head of the household variable does not show much difference. It seems that the gender of the head of household has nothing to do with the household acceptance status program. The gender class of household heads showed that one-third of female households are program recipient households. In the status variable for rural areas, the proportion between the recipient and non-recipient households was not much different. In the marital status variable, the head of household showed that one-third of the household heads with divorced lived/death status as program recipient households.

Identification of Categorical Explanatory Variables Based on Housing

Figure 5 is shown that the proportion of household status shows differences in the explanatory variables of roof type, floor type, wall type, and building status. The stacked bar chart of roof type, floor type, and wall type shows an increasing pattern (shown in the variable category class from left to right). The lower the quality of the roofs, floors, and walls owned by a household, the higher the probability that households will receive the program. In the category class is “others” (the lowest quality) for the variables of roof type, floor type, and wall type; almost 60% are program recipient households. Those included in the "others" category are bamboo, wood/shingle, and straw/ijuk/leaves/rumbia (roof type), logs and bamboo (wall type), and bamboo and soil (floor type). The stacked bar chart for the building status variables shows that more than a quarter of households with their own house is a house program recipient ladder. Another thing that shows that more than a quarter of a household with the status of building a house is free of rent is a house program recipient ladder.

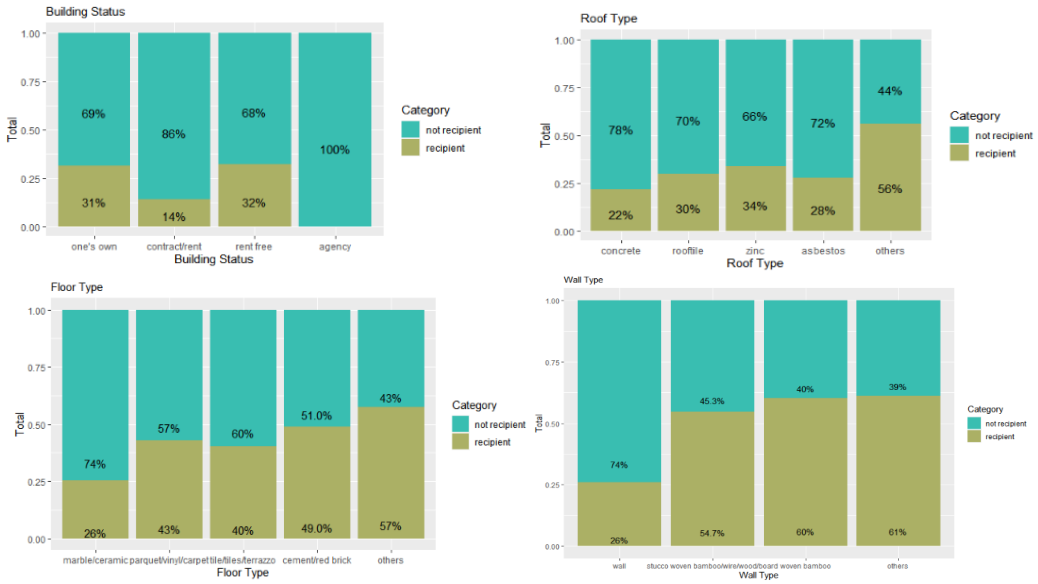

Figure 5. The Proportion of Categorical Variables Based on Housing

Identification of Categorical Explanatory Variables Based on Housing Support

Figure 6 shows that the proportion of household status shows differences in the explanatory variables of defecation facilities, drinking source, light source, and cooking fuel. More than half of the category class that does not have defecation facilities are program recipient households (56%). In the explanatory variable for drinking water sources, it can be seen that the “other” classes (rivers/lakes/reservoirs/ponds/irrigation/ rainwater) have almost the same proportion between recipients and non-recipients of the program. More than half are program recipient households in the category class where the cooking fuel is charcoal/firewood (58%). Stacked bar chart for cooking fuel variable, class category further to the left is the highest quality in the type of fuel used for cooking. Those that fall into other categories are electricity, city gas, biogas, and others.

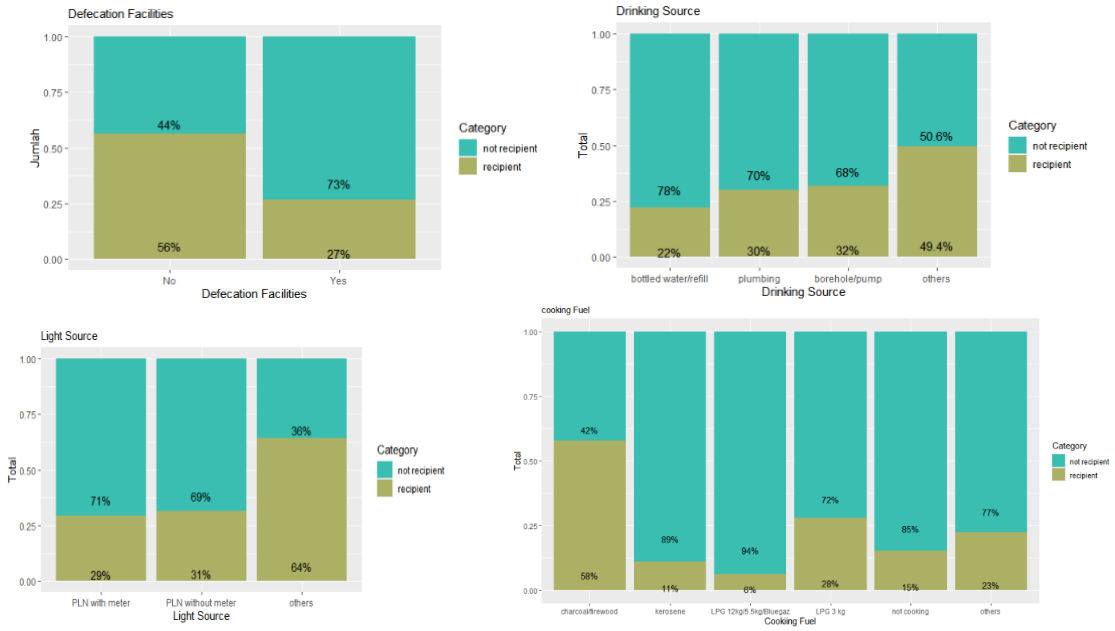

Figure 6. The Proportion of Categorical Variables Based on Housing Support

Identification of Categorical Explanatory Variables Based on Ownership of goods

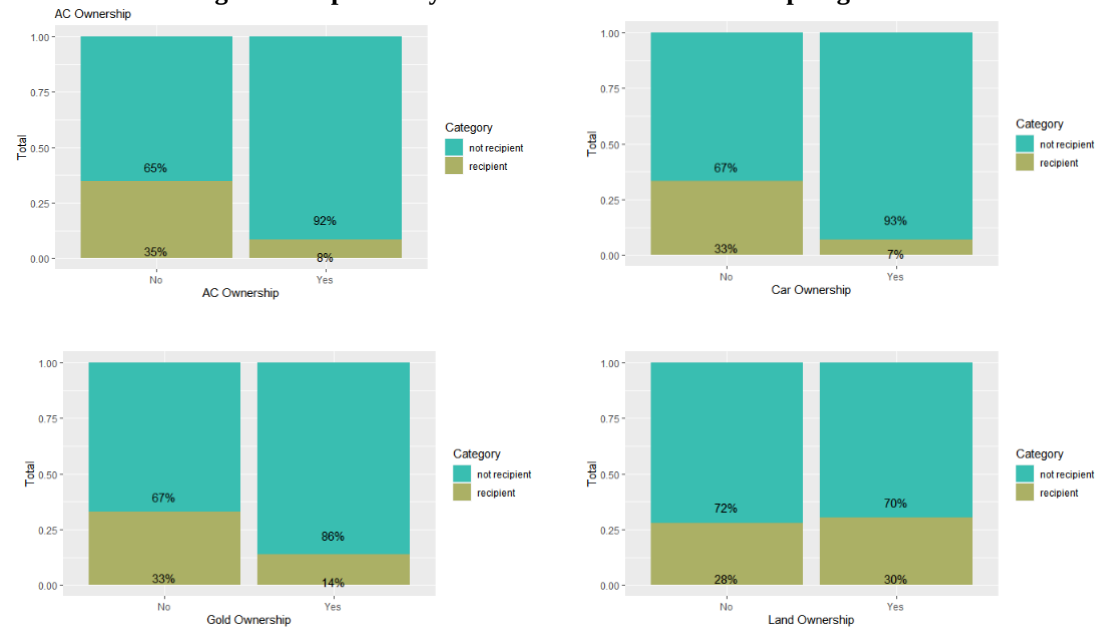

Figure 7. The Proportion of Categorical Variables Based on Ownership of goods

Figure 7 is shown that the proportion of household status shows differences in the explanatory variables of AC, car, and gold ownership. Almost one-third of the households that own land are program recipient households. 110 program recipe air conditioning (AC), 65 program recipient households own a car, 1.616 program recipient households own land, and 165 program recipient households own a minimum of 10 grams of gold/jewelry.

Identification of Categorical Explanatory Variables Based on Education and Employment

Figure 8 is shown that the proportion of household status shows differences in the education level variable. The stacked bar chart of the education level variable shows an increasing pattern (from left to right in the category class). In contrast, the working status variable does not show a significant difference. The stacked bar chart of the education level variable shows that almost half of the households in the primary school class are program recipients (44%), and 46,9% in the no school class are program recipients.

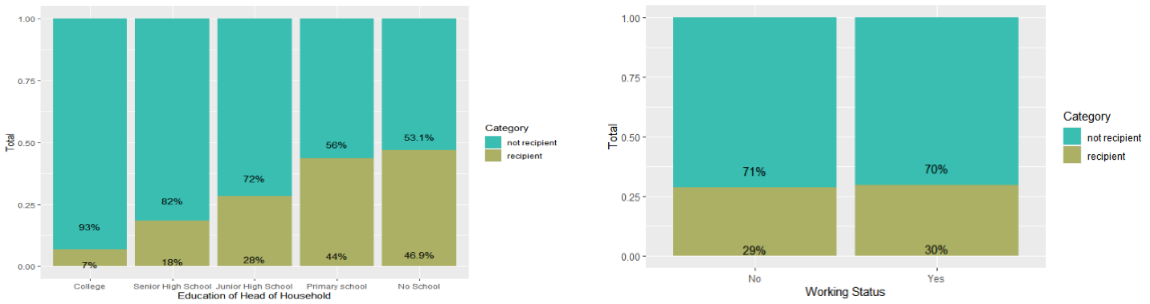

Figure 8. The Proportion of Categorical Variables Based on Education and Employment

3.3. Classification Model with Random Forest

At this stage, a 10-fold cross-validation process is used by considering the optimum hyperparameters in a random forest, namely the number of trees, the number of features considered to find the best split, the maximum depth of the tree, and the minimum number of samples required to be in the leaf node. The results of the 10-fold cross-validation process using training data divided into ten parts with one part as test data and nine other parts being used as training data will produce the optimum value for the hyperparameter. This process uses 100 iterations using training data, and model evaluation is carried out using test data to measure the model's goodness by looking at accuracy, specificity, sensitivity, and AUC. The optimum hyperparameters and the goodness of fit test model are shown in Table 4.

Table 4. Hyperparameter and Classification Model Goodness Measures

Hyperparameter	Value	Goodness Measures	Average
n_estimators	276	Accuracy	0,648
Max_features	3	Sensitivity	0,698
Max_depth	31	Specificity	0,718
Min_samples_split	2	AUC	
Min_samples_leaf	2		
Criterion	Gini		

3.4. Importance Variables with Permutation Feature Importance

The importance variables calculated the model's prediction error increase after permuting the features/variables. A feature is "important" if shuffling its values increases the model error because in this case, the model relied on the feature for the prediction [9]. The scoring of important variables starts from the largest that contributes to the model, as shown in Figure 9. The important variables obtained are capita expenditure, education level of head of household, age of head of household, drinking source, floor area, and total expenditure.

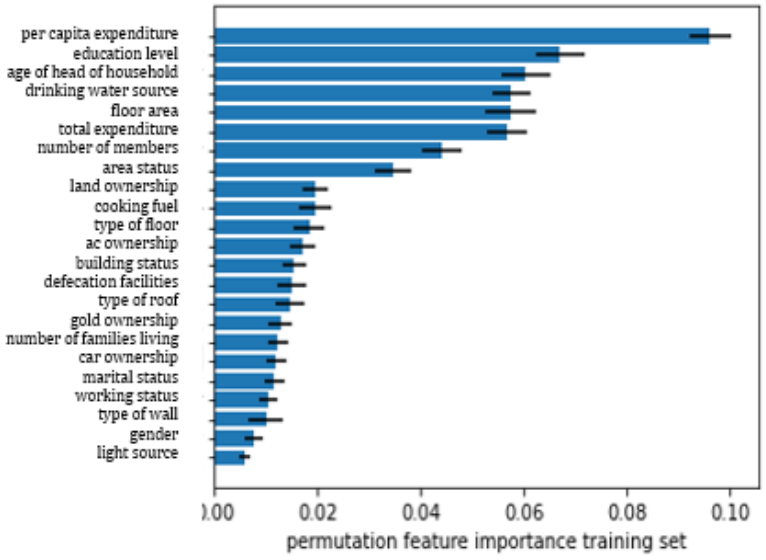

Figure 9. Importance Variables with Permutation Feature Importance

3.5. Importance Variables with SHAP

The measurement of the level of importance of the variables used using SHAP produces the important variable with the highest contribution score in estimating household acceptance of the social protection program based on the resulting Shapley value, as shown in Figure 10. The highest important variable is education level, per capita expenditure, type of wall, AC ownership, drinking source, and age of head of household. The first important variable is the education level of the head of the household. The lower the education of the head of the household, the lower Shapley's score will be. It shows that the education of the head of the household is lower, and the probability of the household accepting social protection programs is higher. The education of the head of the household and the source of drinking water are important variables that have a major contribution to food insecurity [2] [3]. The distribution of red and blue data on the variables of gold ownership, land ownership, type of roof, number of families living, marital status, working status, gender, and light source shows that the red and blue positions are inconsistent in the negative class or positive class. It shows that these variables are unimportant.

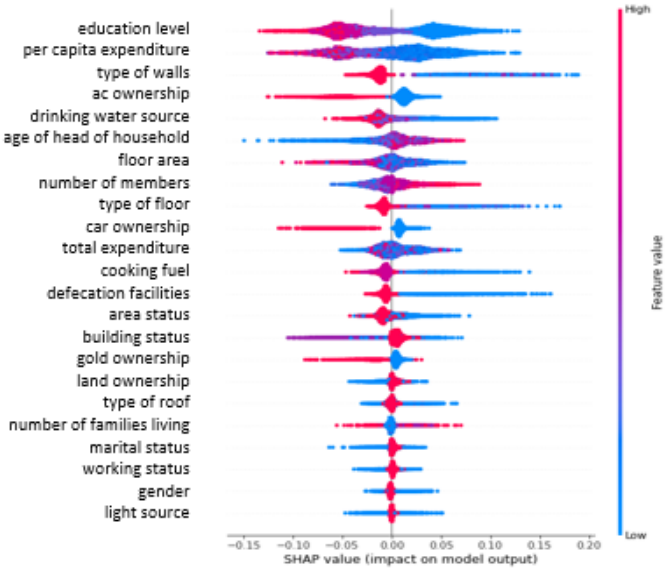

Figure 10. Importance Variables with SHAP

3.6. Biplot Analysis

The important variables obtained from the two techniques, PFI and SHAP, were compared to obtain the final important variables, which would be analyzed further. The important variable is the important variable that has a consistent score/rank at the top of both methods. The important variables were capita expenditure, education level, age of head of household, drinking source, floor area, and number of members. Visualization with a biplot is used to see the relationship between each class category of the six important variables with household status.

The results showed that the random forest technique was an excellent classifier with an AUC value of 0,718. The interpretation techniques of both PFI and SHAP show results that are not much different when ranking variables consistently at the top and bottom ranks. The PFI interpretation technique seems less consistent than the SHAP technique, especially if the variables used have interactions or correlations, so caution is advised in their use. The importance of permuted features can be biased or unrealistic when the features used are correlated [9]. This study found that the correlation of two numeric variables was only capita expenditure and total expenditure had a high enough association (0,788). It was found that there was an association between two categorical variables (a combination of categorical variables).

One of the characteristics of food-insecure households was the education level in Primary School or lower [2], where food-insecure households are the target of providing social protection programs. The results of this study also show that education level (maximum of Primary School/equal) was a characteristic of the program recipient households [3].

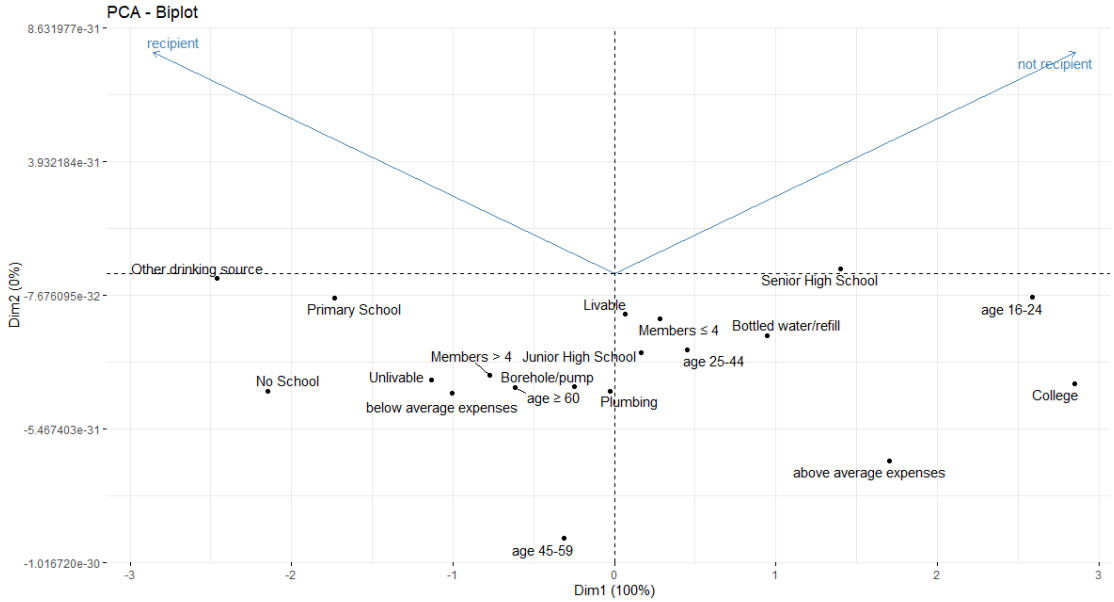

Figure 11. Factor Map of Household Variable Categories and Household Status

5. CONCLUSION

The permutation feature importance technique is less consistent than the SHAP technique. Based on the important variables produced by the permutation feature importance and SHAP methods, most of them were not much different from ranking variables at the top and lowest rankings. The results showed that the characteristics of the households receiving the program were other drinking sources, namely protected wells, unprotected wells, protected springs, unprotected springs, surface water (rivers/lakes/reservoirs/ponds/irrigation/rainwater), education level of head of the household was maximum of Primary School, capita expenditure was below the average per capita expenditure of Banten Province in 2021, the floor area was unlivable, several members > 4 peoples, the minimum age for the head of the household was 60 years.

REFERENCES

[1] Muhamad FI. Pengaruh Bantuan Sosial terhadap Kemiskinan di Indonesia. Undergraduare Thesis. Bogor: IPB University; 2021.
[2] Irawan H. Faktor-Faktor Rumah Tangga yang Mencirikan Tingkat Kerawanan Pangan. Master Thesis. Bogor: Postgraduate IPB University; 2019.
[3] Irfani R. 2021. Pendekatan Eksploratif untuk Melihat Peubah Penciri Rumah Tangga Berdasarkan Kemiskinan dan Kerawanan Pangan. Master Thesis. Bogor: Postgraduate IPB University; 2021.
[4] Suswanto D. Analisis Perbandingan Metode *Machine Learning* pada Prediksi Khasiat Jamu. Undergraduare Thesis. Bogor: IPB University; 2016.
[5] Farel F. Pemodelan Klasifikasi Keterlambatan Pembayaran UKT Mahasiswa IPB dengan Random Forest dan AdaBoost. Undergraduare Thesis. Bogor: IPB University; 2021.
[6] Albasia MAY. Klasifikasi Keberhasilan Melanjutkan Pendidikan Jenjang SMA di Provinsi Banten dengan Metode CART dan Random Forest. Undergraduare Thesis. Bogor: IPB University; 2018.
[7] Rosita, AA. Evaluasi Metode *Ensemble* untuk Klasifikasi Multi Kelas Data Tak Seimbang. Undergraduare Thesis. Bogor: IPB University; 2021.
[8] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research. 16(1):321-357.
[9] Molnar, C. Interpretable Machine Learning. A Guide for Making Black Box Models Explanaible. https://christophm.github.io/. 2021.
[10] Breiman L, Friedman J, Stone C, Olshen R. Classification and Regression Trees (Wadsworth Statistics/Probability). New York CRC Press. 1984.
[11] Agresti A. Categorical Data Analysis. New Jersey (US); Wiley. 2002.
[12] Lundbergunberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. Volume ke-2017-Desember: 4768-4777, arXiv: 1705.07874v2. 2017.
[13] Burnaev E, Erofeey P, Papanov A. Influence of Resampling on Accuracy of Imbalanced Classification. *Eighth International Conference on Machine Vision* (ICMV). 9875. 2015.
[14] Bappenas (Badan Perencanaan Pembangunan Nasional). Perlindungan Sosial di Indonesia: Tantangan dan Arah ke Depan. Cetakan I. Jakarta; Bappenas. 2014.
[15] BPS (Badan Pusat Statistik). Profil Kemiskinan di Indonesia. Jakarta; BPS RI. 2021.

[16] Breiman L. Random Forest. Mach Learn. 45(1):5-32. https://doi.org/10/1023/A:1010933404324. 2001.
[17] James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning with Applications in R. Springer; Network. DOI: https://doi.org/10.1007/978-1-4614-7138-7. 2013.
[18] Katsuya F, Fukazawa Y, Kapoor N, Kito T. Pairwise acquisition prediction with SHAP value interpretation. *The Journal of Finance and Data Science 7* (2021) 22-24. 2021.
[19] Marie AG, Kaufmann T, Quintana DS, Winterton A, Andreassen OA, Westlye LT, Ebmeier KP. Prominent health problems, socioeconomic deprivation, and higher brain age in lonely and isolated individuals: A population-based study. DOI: 10.1016/j.bbr.2021.113510. 2021.
[20] Mattjik AA, Sumertajaya IM. Sidik Peubah Ganda dengan Menggunakan SAS. Bogor; IPB University. 2011.
[21] Raquel R, Bajorath J. Chemistry-centric Explanation of Machine Learning Models, Artificial Intelligence in the Life Sciences. DOI: https://doi.org/10.1016/j.ailsci.2021.100009. 2021.
[22] de Lange, A. G., Kaufmann, T., Quintana, D.S., Winterton, A., Andreassen, O.A., Westlye, L.T., et al. Prominent Health Problems, Socioeconomic Deprivation, and Higher Brain Age in Lonely and Isolated Individuals: A population-Based Study. UK; University of Oxford. 2021. Doi:10.1016/j.bbr.2021.113510
[23] Hidayati, E.F.K. Metode Global Surrogate dan Shapley Additive Explanations (SHAP) untuk Menjelaskan Model Klasifikasi Industri Pelaku Litbang. Master Thesis. Bogor: Postgraduate IPB University; 2021.

BIBLIOGRAPHY OF AUTHORS

Nofrida Elly Zendrato. Statistician and coordinator of sub-district statistics at BPS. Formal educational background in mathematics and statistics obtained from the University of North Sumatera and IPB University. Work experience in the technical field in several surveys in social, production, distribution, integration of statistical processing and dissemination, and regional accounting and statistical analysis.

Bagus Sartono. Lecturer at IPB University. Formal education background in statistics from IPB University and Applied Economics from Universiteit Antwerpen. His field of expertise are data mining and machine learning.

Utami Dyah Syafitri is a lecturer in the Department of Statistics, Faculty of Mathematics and Natural Science, IPB University. Her bachelor's and master's were graduated from IPB University, Indonesia. She did a Ph.D. at Antwerp University, Belgium. Her expertise are in experimental design, optimal design, classification, and modeling.