ICDSOS
International Conference
on Data Science and Official Statistics

"Harnessing Innovation in Data Science and Official Statistics to Address Global Challenges towards the Sustainable Development Goals"

Implementation of Machine Learning and Its Interpretation for Mapping Social Welfare Policy in Indonesia

A L Irfiansyah[1,*], **A Rismansyah**[1], **N Permatasari**[1], **I Noviyanti**[1], **A Mardiyanto**[1], **A Koswara**[1]

[1]BPS-Statistics Indonesia, Jakarta, Indonesia

*Corresponding author's email: aldo.leofiro@bps.go.id

Abstract. This research leverages data from the 2022 Early Socio-Economic Registration (Regsosek) activity to develop a machine learning model capable of predicting family expenditure levels based on the Proxy Mean Test (PMT) with high accuracy. By integrating the SHAP (SHapley Additive exPlanations) method for model interpretation, we identify the contributions of socio-economic features to expenditure predictions and link them to relevant social assistance programs. We compare two regions, Kulonprogo Regency and Yogyakarta City, representing varying poverty levels, and identify unique characteristics influencing family welfare in each area. The results highlight that effective policy interventions must be tailored to the unique characteristics of each region and family, taking into account dimensions such as housing, education, income, and community expenditures. This research provides valuable insights for policymakers, demonstrating that successful poverty alleviation policies are data-driven and adaptable to the diverse socio-economic realities across regions.

1. Introduction

Eradicating poverty in all its forms and dimensions constitutes the greatest global challenge and a crucial prerequisite for sustainable development [1]. Countries worldwide have committed to eradicate poverty, stated as the primary goal of Sustainable Development Goals, Goal 1: "End poverty in all its forms everywhere"[2]. In Indonesia, poverty is also being a priority as stated in the National Medium-Term Development Plan for 2020-2024 that the poverty rate is targeted to decrease to 6.0 – 7.0 percent by 2024 (*Appendix Presidential Regulation No 18 of 2020*).

To measure poverty, BPS-Statistics Indonesia employs the basic needs approach, which assesses the economic incapability to fulfill both food and non-food basic necessities, measured by family expenditure [3]. According to BPS data, in March 2023, the poverty rate in Indonesia remained at 9.36%, slightly higher than the pre-pandemic condition in the second semester of 2019. The years 2020-2021 showed a significant impact of COVID-19 on the Indonesian socio-economy condition, which led to a significant increase in Indonesia's poverty rate to 10.19% in the second half of 2020. As of the first semester of 2023, Indonesia's poverty trend shows a declining pattern, aiming to reach the target of 6-7% by 2024.

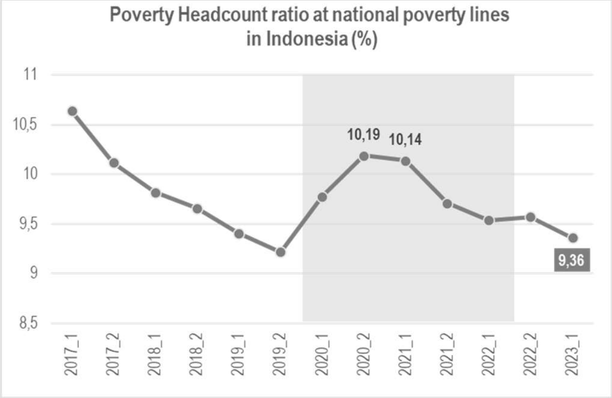

Figure 1. Poverty headcount ratio at national poverty line in Indonesia, 2017-2023.
(Source: BPS-Statistics Indonesia)

The effective and efficient poverty eradication program necessitates the involvement of multiple stakeholders, including the government, private sector, broader society, and the community themselves [4]. The government, as the most responsible entity in addressing poverty issues, is expected to demonstrate a strong commitment to implementing poverty eradication strategies [5]. In Indonesia, poverty alleviation programs are categorized into two main types: (1) the program delivery approach and emergency interventions, which are characterized by their urgent and short-term nature; and (2) capacity building and sustainability programs, which represent long-term strategic initiatives focused on enhancing community capacity and empowerment [4]. One of the short-term programs implemented by the government involves the distribution of social assistance to alleviate the financial burden on families, particularly those in poverty [6].

In the implementation, social welfare initiatives in several areas have not been effective [7][8][9]. This is due to inaccuracies in targeting beneficiaries, unequal distribution of aid, and various frauds in the distribution of social assistance [8]. Improving the accuracy of beneficiary targeting requires an accurate and up-to-date database.

In 2022, BPS conducted a Social Economic Registration (Regsosek) to record the socio-economic conditions of the Indonesians. Variables collected include population and employment, social protection, housing area, education, disability health, and economic empowerment [10]. The data collection unit is 100% of the population in Indonesia.

The government in collaboration with the National Team for Accelerated Poverty Reduction (TNP2K) uses the Proxy Means Testing (PMT) method to target the poor [11][12]. Proxy Means Testing utilizes information about family or individual characteristics that correlate with the level of well-being to represent family income, welfare, or needs [13]. Proxy means testing uses a regression model to predict the expenditure.

Data collected by Regsosek on individuals' names and addresses can be used to provide more targeted social assistance. The socio-economic characteristics of families can be used to determine which program is most suitable. A machine learning model can be used to identify the most important characteristics. The importance of variables in a machine learning model can be further interpreted by understanding which variables have the greatest influence on the model. In this research, we want to use the model interpretation for more accurate beneficiary targeting by using Regsosek by-name-by-address data.

2. Study Area

The locus of this research is in Yogyakarta Province, Indonesia. Based on BPS-Statistics Indonesia data, Yogyakarta has the highest percentage of poor people on the island of Java, 11,04 percent in March 2023. Yogyakarta is the highest Gini coefficient province in Indonesia. A Gini coefficient of 0.449 indicates that income inequality is high. In 2022, the district with the highest poverty rate in the Special Region of Yogyakarta is Kulonprogo, with a poverty rate of 16,39%, and the district with the lowest poverty rate is Yogyakarta City, with a poverty rate of 6,62%. Kulonprogo is a rural regency with a relatively low level of development that makes it more difficult for residents to access jobs, education, and other opportunities. Meanwhile, Yogyakarta City is a major urban center with a strong economy, a major center for tourism, and a relatively high level of education.

The high poverty rates in Kulonprogo Regency and the low poverty rates in Yogyakarta City provide an interesting opportunity for further research on the factors that influence poverty in the region. A comparison of the influencing factors in these two districts could help to identify effective strategies for overcoming poverty by better and more accurate beneficiary targeting.

3. Literature Review

3.1. Proxy Mean Test

A proxy means test is a method of targeting social programs for the poor by using a set of family characteristics that are correlated with income. These characteristics can include things like the number of people in the family, the type of housing, the ownership of assets, and the level of education. By using statistical methods to identify the family characteristics that are most strongly linked to income, create a scoring system to indicate higher/lower income. If the family's score is below a certain threshold, it is eligible for the program.

3.2. Classification Model

Various machine learning models have been investigated to better target the poor, as shown in the literature review in Table 1.

Table 1. Literature Review.

Study Area	Method	Result	References
Indonesia	MARS, K-Nearest Neighbor, Decision Tree, and Bagging	Machine learning models were slightly more accurate than PMT models in predicting family socioeconomic status, with MARS being the best-performing model.	(Taufiq and Mariyah 2021) [14]
Ghana	ML-based PMT	Our field assessment found that the new ML-based PMT was more effective than other approaches to identifying families having assets associated with wealth.	(Poulin et al. 2022) [15]
Togo	LightGBM	The machine learning approach was more accurate than geographic targeting options in identifying families eligible for assistance in Togo, reducing the error rate of exclusion.	(Aiken et al. 2022) [16]
Thailand	Random Forest	The PMT model based on variable selection of RF is more effective than other PMT models in minimizing exclusion errors, so it is more appropriate for social welfare programs.	(Kambuya 2020) [17]

Several machine learning studies also show that Catboost is a powerful classification method, especially those involving categorical and heterogeneous data [18]. Considering the models that have been used in previous studies and the costs of processing, we compare several models: decision tree, random forest, gradient boosting, and Catboost.

3.2.1. Decision Tree. Decision trees are fundamental machine learning models that recursively partition data based on feature conditions. Each split represents a decision path leading to a final prediction. Decision trees are known for their interpretability and are widely used in various applications [19]. However, they may suffer from overfitting issues, which led to the development of ensemble methods.

3.2.2. Random Forest. Random Forest is a robust ensemble learning technique that extends the capabilities of decision trees to improve predictive accuracy and reduce overfitting [20]. By constructing multiple decision trees with random subsets of training data and features, Random Forest introduces diversity and stability, making it adept at handling noisy and high-dimensional datasets. Its ensemble approach, which aggregates predictions from individual trees, leads to more accurate overall forecasts. Furthermore, Random Forest provides valuable feature importance scores, aiding in identifying the most influential variables in predictive modeling tasks. Therefore, it is widely applied in various domains, offering a powerful tool for complex classification and regression tasks.

3.2.3. Gradient Boosting. Gradient boosting is a powerful ensemble learning technique that iteratively combines weak predictive models, often in the form of decision trees, to create a robust overall model [21]. It focuses sequentially on instances where prior models performed poorly, refining predictions with each iteration. Gradient boosting has been highly regarded for its capacity to enhance model accuracy and manage overfitting, making it a valuable tool across various applications in machine learning.

3.2.4. Catboost. CatBoost, an advancement in gradient boosting, specifically addresses categorical feature management. CatBoost introduces an "ordered boosting" technique that integrates categorical hierarchy, refining predictive accuracy [22]. By considering categorical variable order, CatBoost optimizes boosting, resulting in robust generalization. The algorithm adeptly handles missing values during tree construction, streamlining preprocessing. These attributes collectively establish CatBoost as an efficient and precise choice for predictive modeling tasks involving categorical attributes [23].

3.3. Evaluation Metrics

3.3.1. K-Fold Cross-Validation. K-Fold cross-validation is a fundamental technique for assessing machine learning model performance. It involves splitting the dataset into "k" subsets or folds, using "k-1" folds for training and the remaining fold for testing. This process is repeated "k" times, with each fold serving as the test set once. K-Fold cross-validation provides a more robust estimate of a model's generalization performance by mitigating the risk of overfitting or relying on a single train-test split [24]. An illustration of k-fold cross-validation is shown in the following:

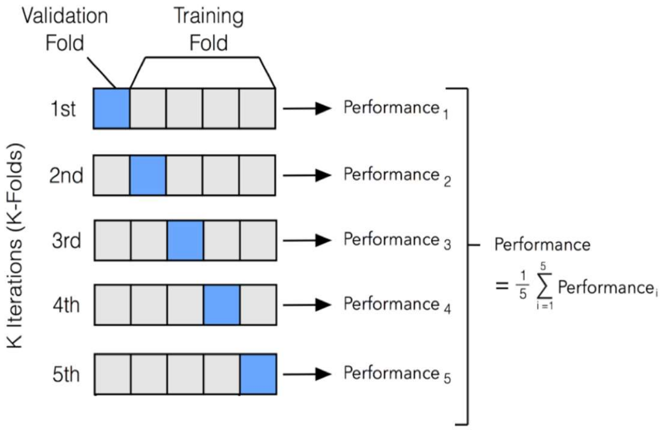

Figure 2. K-fold cross-validation illustration

3.3.2. Mean Absolute Percentage Error (MAPE). MAPE is an evaluation metric that quantifies the average percentage difference between predicted and actual values. It is particularly useful for measuring the accuracy of models in predicting continuous values, such as in regression tasks. The formula for MAPE is:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{1}$$

Where y represents the actual value, ŷ is the predicted value, and *n* is the numbers of instances.

3.3.3. Coefficient of Determination (R2). R2, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, where 1 indicates perfect prediction. R2 assesses how well the model's predictions align with the actual variability in the data. It is computed as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \underline{y_i})^2} \tag{2}$$

Where y represents the actual value, ŷ is the predicted value, ȳ is the mean of the actual values and *n* is the numbers of instances.

In conclusion for choosing the best model, K-Fold cross-validation enhances model assessment by minimizing overfitting risks. MAPE and R2 serve as evaluation metrics for regression tasks, capturing prediction accuracy and variability explained by the model, respectively.

3.4. *SHAP Value*

SHAP (SHapley Additive exPlanations) values provide a method to interpret the output of machine learning models. They offer insights into the contribution of individual features towards model predictions. SHAP values are based on the cooperative game theory concept of Shapley values, which fairly allocate the value generated by coalition players [25]. In the context of machine learning, SHAP values attribute specific contributions of each feature to deviations in the model's predictions from the expected outcome and these contributions collectively explain the final prediction.

The SHAP value for a specific feature *i* for a given instance *x i*s computed as the difference between the model's prediction for *x* and the average prediction over all possible coalitions that exclude feature *i* where *S* represents the set of features:

$$SHAP_i(x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[\hat{f}_S(x) - \hat{f}_{S \setminus \{i\}}(x) \right] \tag{3}$$

Where:

- *N* is the set of all features
- *S* is a subset of features excluding feature *i*
- |*S*| is the size of subset *S*
- fs(*x*) is the model's prediction for instance when features *x* in subset *S* is active

This formula calculates the average difference in predictions caused by including feature *i* in the model, considering all possible combinations of features.

Implementing SHAP values involves decomposing predictions into contributions from individual features, collectively summing to the final prediction. These contributions provide a clear view of how each feature influences the overall prediction. Larger absolute SHAP values indicate more substantial impacts, facilitating the identification of crucial drivers within the model.

4. Methodology

4.1. *Data Description*
The dataset used in this machine learning model is derived from the Early Social Economic Registration (Regsosek) activity of 2022. Through Regsosek, comprehensive data is collected encompassing individual profiles, socio-economic conditions, and welfare levels within each family. These attributes will be further processed and then utilized as a feature (X) in the machine learning modeling process. The model's goal is to learn patterns and relationships between these features and the estimated expenditure levels derived from the Proxy Mean Test (PMT).

By focusing solely on Regsosek data, the model is designed to offer policy recommendations based on the socio-economic characteristics of each family, aiming to predict their expenditure levels. This approach aids decision-making processes by leveraging insights from the gathered data to provide targeted and informed recommendations. The model's predictions could have significant implications for policy planning and resource allocation, ultimately contributing to more effective socio-economic initiatives.

The dataset from the 2022 Regsosek registration serves as the foundation for the machine learning model, utilizing socio-economic attributes as features and PMT-derived expenditure estimates as labels. This targeted approach enables the model to provide relevant and tailored policy suggestions centered around expenditure predictions.

4.2. *Modeling Pipelines*
The research methodology incorporates a pipeline modeling approach to ensure the systematic and efficient development of predictive models. The research implements modeling pipelines to construct, evaluate, and interpret machine learning models, as shown in Figure 3.

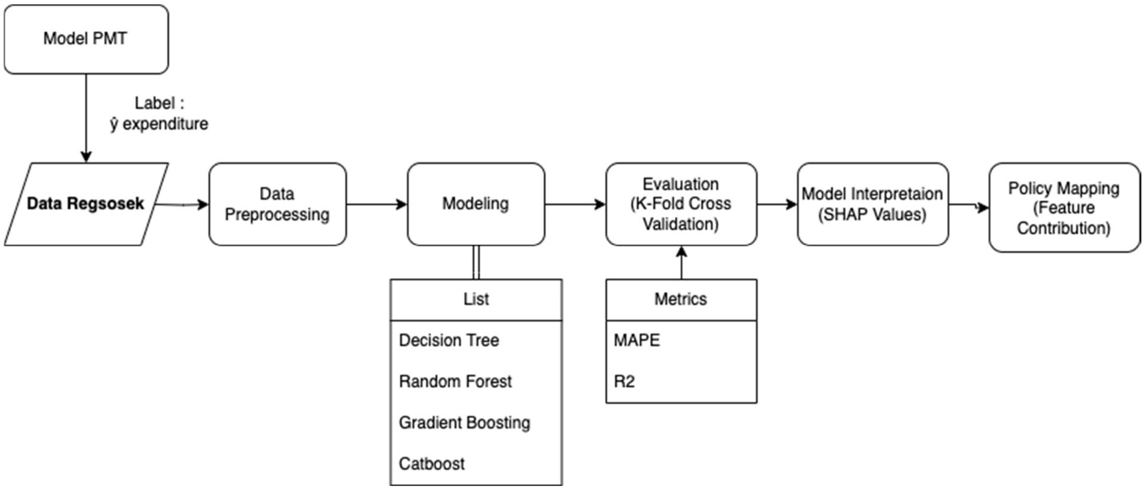

Figure 3. Modeling pipeline.

These pipelines encompass a series of interconnected stages that facilitate the systematic progression from raw data to insightful model outcomes. The process commences with data preprocessing, encompassing tasks such as data cleaning, feature engineering, aggregation, joining, and encoding categorical features. This initial phase ensures the data is appropriately prepared for subsequent modeling steps.

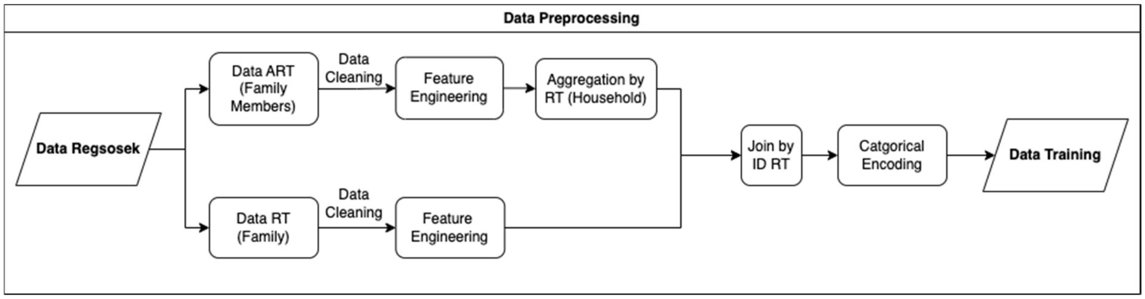

Figure 4. Data preprocessing pipeline.

Following data preprocessing, shown in Figure 4, the pipeline advances to model selection and training. Diverse algorithms, including Decision Trees, Random Forest, Gradient Boosting, and CatBoost, are implemented and fine-tuned through cross-validation techniques to optimize their performance. The evaluation stage involves rigorous assessment using K-Fold cross-validation, where evaluation metrics like *Mean Absolute Percentage Error* (MAPE) and *Coefficient of Determination* (R2) gauge model accuracy and predictive power.

The next stage of the pipeline involves interpreting model predictions through SHAP (SHapley Additive exPlanations) values. SHAP values provide a nuanced understanding of feature contributions to individual predictions, guiding insights into the model's decision-making process. This enables researchers to map these insights to relevant policies, enriching the practical utility of the model's recommendations.

This holistic approach ensures that the entire modeling process is transparent, coherent, and adaptable. The pipeline encapsulates the entire lifecycle, from data preparation to model interpretation, within a single framework. By systematically guiding the researcher through each step, pipeline modeling guarantees methodological rigor and enhances the applicability of the research outcomes to real-world policy decisions. Through this approach, the research contributes not only to predictive accuracy but also to the meaningful interpretation and translation of model insights into actionable policies.

4.3. Model Interpretation

The interpretation of the generated machine learning model is facilitated by leveraging the SHAP (SHapley Additive exPlanations) value methodology, enabling a comprehensive understanding of feature contributions to individual predictions. SHAP values differ from traditional feature importance metrics and feature correlation analyses by offering a more detailed explanation of how individual features influence machine learning model predictions. While traditional metrics may provide the magnitude of feature importance and explore associations between features, SHAP values concentrate on explaining how each feature's presence or absence influences the model's outputs and the direction (positive or negative) of a feature's impact on predictions. In this context, SHAP values offer a valuable, granular tool for model interpretation by revealing the specific contributions of each feature to model predictions, enhancing our comprehension of model behavior. This approach aligns with the study's objective of mapping these feature contributions to pertinent policies, enhancing the practical utility of the model's insights.

Regarding policy mapping, SHAP values offer a bridge between model insights and actionable policies. Features with significant positive or negative SHAP values spotlight attributes with pronounced influences on predictions. By linking influential features to distinct policy domains, decision-makers can harmonize interventions with socio-economic context. For example, if an education-related feature exhibits a notable positive SHAP value, it suggests enhancing educational policies might yield favorable effects on the model's predicted outcomes.

SHAP values provide an interpretable framework for comprehending individual feature contributions in machine learning models. By translating these insights into actionable policies, decision-makers can utilize the model's output to formulate targeted interventions. This integration ensures that policy recommendations draw from the model's learned relationships, ultimately enhancing the effectiveness of socio-economic initiatives.

It is important to note that SHAP values are model-specific and may differ if calculated from other machine learning models, such as Random Forest (RF), Decision Trees (DT), or Gradient Boosting (GB). SHAP values offer a unique and detailed perspective on feature contributions within the context of a specific model. While the principles of SHAP values are consistent, the actual values may vary based on the underlying model's structure and prediction behavior. In our analysis, all SHAP values presented are derived from the best algorithm model, which is selected later based on performance metrics. Therefore, our SHAP values are specific to the selected best model, and this specificity should be considered when interpreting the results. It's essential to understand that SHAP values represent the contribution of individual features to predictions made by the selected model and similar contributions may not be obtained when using other machine learning algorithms.

4.4. Policy Mapping

Policy mapping is conducted based on contributions of features in the form of percentages, both at the family and regional levels. To formulate program recommendations, an examination of the suitability is carried out between features that contribute to family welfare with criteria of each social protection and community empowerment program. There are several programs that are stated in the Decree of the Coordinating Minister for Human Development and Culture of the Republic of Indonesia Number 32 Year 2022. This research will show program recommendations that have a high percentage of matches based on the comparison between features and criteria. The obtained results can be used to identify which interventions should be undertaken to enhance family welfare based on the high contribution of features at the family level. In addition, the results can also be used to identify programs that are in line with average value of feature contribution at the regional level.

The features are sorted based on contribution value (average of percentage of absolute SHAP Value) from the largest to the smallest. Then for each criteria of social protection and community empowerment program a number of features with the largest contribution are selected with the same amount of each criterion. The percentage of features that match with the criteria of each social protection and community empowerment program used as the basis for making recommendations.

$$Percentage\ of\ Match = 100\% \times \frac{n_{match}}{n_{criteria}}$$

5. Result

5.1. Modeling Result

In this section, we compare the performance of four machine learning models to determine the best model for predicting expenditure estimation by the Proxy Mean Test model. The results of the different regression algorithms are shown in Table 2.

Table 2. Comparison of model's performance measures.

Models	*R2*	*MAPE (%)*
Gradient Boosting	0.529362	27.2308
Decision Tree	0.411651	30.7984
Random Forest	0.435157	30.6099
Catboost	0.71296	20.7652

Based on model performance as measured by R2 and MAPE, Catboost algorithm has better performance than the other three algorithms. Therefore in this study, we choose the Catboost algorithm to predict estimated expenditure levels derived from the Proxy Mean Test.

5.2. Model Interpretation using SHAP

Prediction results from the model with the Catboost algorithm interpreted by SHAP value to provide insight into how each feature contributes to the estimated expenditure that is derived from the proxy mean test. Percentage of the absolute SHAP value used to see more clearly how it contributes to the estimated expenditure for each family. More broadly we can see the characteristics of families in the same region by using the aggregate value which is the average of the percentage of absolute SHAP value of each feature/variable.

Table 3. List of 20 Features with the Largest Contribution (Average of Percentage of Absolute SHAP Value) in Kulonprogo Regency.

Variables Name	Average of Value SHAP	Average of Value SHAP Percentage
h_tfloor	-1345.72	12.05
h_hhcount	-376.13	9.11
h_pcfloor	1630.32	8.58
h_cookingfuel	-768.31	7.37
h_ngrad_s	-1068.44	5.33
h_dwater	738.22	4.57
h_kk_pendidikan_grad_sma	-510.66	3.82
h_kk_pendidikan_grad_s	1032.53	3.23
h_toiltype	-290.48	2.97
h_nfemale	341.4	2.81
h_njamkerja	-146.39	2.7
h_ngrad_sma	1079.9	2.26
h_nage0519	-144.44	2.06
h_nageod	79.36	1.84
h_akses_internet	-433.1	1.58
h_house	-226.04	1.41
h_nmale	-152.04	1.4
h_ngrad_d	267.32	1.37
h_kk_pendidikan_grad_d	135.18	0.94
h_nfemale_age0519	-159.2625	0.092

In Table 3 there are twenty features with the largest value of average percentage SHAP value for families in Kulonprogo regency. It can be seen that the largest average percentage of SHAP value is dominated by features that are related to housing and education such as the widest type of floor. the floor area of the building residence per capita. fuel/main energy for cooking, main drinking water source, and the education level of the head and family members. Table 4 shows the one sample of families in

Kulonprogo regency that has similar characteristics to those described by the aggregate value (average of percentage SHAP value).

Table 4. List of 20 Features with Largest Contribution (Percentage of Absolute SHAP Value) of family in Kulonprogo Regency.

Variables Name	Value of Variable	Value of SHAP	Value of Percentange SHAP	Dimension
h_tfloor	1.0	1127934.13	10.397616	Housing
h_dwater	3.0	1-114192.74	9. 280809	Housing
h_hhcount	2.0	103040.43	8.374425	Social
h_kk_pendidikan_grad_s	0.0	-89179.37	7.2478924	Education
h_akses_internet	0.0	-83102.15	6.7539773	Economy
h_ngrad_s	0.0	-57385.46	4.6639	Education
h_toiltype	1.0	55109.496	4.4789248	Perumahan
h_kk_pendidikan_grad_sma	1.0	50496.49	4.104011	Education
h_nage0519	0.0	46642.5	3.790785	Social
h_pcfloor	36.0	35897.066	2.9174693	Housing
h_nonzet_usaha_ultramikro	1.0	-33860.695	2. 7519667	Economy
h_cookingfuel	2.0	33765.816	2.7442558	Housing
h_house	1.0	28564.455	2.3215246	Housing
h_ngrad_sma	2.0	27691.18	2. 2505507	Education
h_sec3_stat6	2.0	125129.635	2. 0423658	Economy
h_nfemale	1.0	21821.07	1.7734683	Social
h_nkepemilikan_izin	0.0	-21266.707	1. 7284133	Economy.
h_njamkerja	42.0	20136.049	1.6365211	Economy
h_nmale	1.0	19255.7	1.5649722	Social
h_sec3_stat2	2.0	115786.286	1.2830019	Economy

For families in Yogyakarta city, in aggregate there are several different characteristics compared to the Kulonprogo regency. In Table 5 there are several features related to socio-economic status that appeared in Yogyakarta but not in Kulonprogo, such as business license ownership, and the number of family members aged 20-64 years. It’s also what resembles the characteristics of one of the families in Yogyakarta city (in Table 6).

Table 5. List of 20 Features with Largest Contribution (Average of Percentage of Absolute SHAP Value) in Yogyakarta City.

Variables Name	Average of Value SHAP	Average of Value SHAP Percentage
h_hhcount	-5395.7297	14.9962
h_pcfloor	16540.9781	12.0584
h_dwater	866.8248	9.1262
h_kk_pendidikan_grad_s	9025.2977	7.4572
h_tfloor	-2930.1414	5.4684
h_ngrad_s	5745.9012	5.4429
h_house	-213.7177	5.2419
h_nkepemilikan_izin	-2377.5745	2.8240
h_kk_pendidikan_grad_sma	-5052.4744	2.7337
h_toiltype	23.2790	2.5104
h_nage2064	1516.9131	2.5053
h_nfemale	-960.3943	2.2169
h_kk_pendidikan_grad_d	-895.3777	1.7585
h_nmale	-816.7442	1.7582
h_ngrad_sma	-9597.0264	1.6863
h_njamkerja	-552.7054	1.6637
h_nage0519	-1641.7753	1.5001
h_akses_internet	-408.9145	1.3977
h_septic	243.0669	1.2075
h_npendidikan_sma	-726.6239	1.0581

Table 6. List of 20 Features with Largest Contribution (Percentage of Absolute SHAP Value) of Family in Yogyakarta City.

Variables Name	Value of Variable	Value of SHAP	Value of Percentange SHAP	Dimension
h_hhcount	4.0	-451487.6	25.280565	Social
h_pcfloor	8.75	-356208.28	19.945501	Housing
Ih_house	1.0	123540.18	6.9175005	Housing
h_dwater	2.0	-108357.02	6.067336	Housing
h_nage0519	1.0	-72581.914	4.0641465	Social
h_nfemale	2.0	-54294.57	3.401664	Social
h_nkepemilikan_izin	0.0	-47647.63	2.667978	Economy
h_nage2064	3.0	-45622.633	2.5545907	Social
h_nmale	2.0	-40628.777	2.274965	Social
h_toiltype	1.0	39784.87	2.2277114	Housing
h_kk_pendidikan_grad_sma	1.0	38272.695	2.1430387	Education
h_ngrad_sna	4.0	34855.812	1.951714	Education
Ih_twall	1.0	33959.62	1.9015326	Housing
h_tloor	2.0	-33130.336	1.8550978	Housing
h_kk_pendidikan_grad_s	0.0	-32590.613	1.8248767	Education
h_ngrad_s	0.0	-31973.33	1.7903125	Education
h_nfemale_age0519	1.0	-20908.41	1.1707441	Social
h_nmale_age2064	2.0	-20526.525	1.1493609	Social
h_nomzet_usaha_ultramikro	0.0	14672.68	0.8215810:	Economy.
h_kk_pendidikan_grad_d	0.0	-13988.815	0.7832888	Education

5.3. SHAP Value for Policy Mapping

The contribution of the SHAP value of each feature can describe the characteristics of a family individually even in aggregate based on region. Mapping the characteristics of families with the criteria of social assistance programs or policies can provide more effective recommendations. The following are recommendations for social assistance programs based on the compatibility of family's characteristics with the criteria of existing social assistance programs expressed in match percentage values. We can see on Table 7 the recommendations of social assistance programs for families in Kulonprogo Regency in general are social rehabilitation of uninhabitable houses, subsidy for 3-kilogram LPG, and Family Hope Program (*Program Keluarga Harapan (PKH)).*

Table 7. Policy Mapping Based on Contributions of Features at Regional Levels in Kulonprogo City.

Dimension	Program	Percentage of Match
Housing	Social rehabilitation of uninhabitable houses	33.333332
Economy	Subsidy for 3-kilogram LPG	14.285714
Housing	Subsidy for 3-kilogram LPG	14.285714
Education	Family Hope Program	6.25
Health	Family Hope Program	6.25
Social	Family Hope Program	6.25

The recommendations can also be unique for each family individually, recommendation for each family derived based on its characteristics. For example. in Table 8, the recommendations for social assistance for the sample family in Kulonprogo regency are Community-based sanitation, Social rehabilitation of uninhabitable houses, Pre-Employment Cards, and Cash assistance to street vendors, stalls, and fishermen. The result recommendations may be different from the recommendations for the region. So that the targeting of aid to each family can be more accurate based on its characteristics.

Table 8. Sample of Policy Mapping Based on Contributions of Features at Family in Kulonprogo City.

Dimension	Program	Percentage of Match
Housing	Community-based sanitation	50.0
Housing	Social rehabilitation of uninhabitable houses	33.333
Social	Pre-Employment Card	11.111
Economy	Pre-Employment Card	11.111
Economy	Cash assistance to street vendors, stalls, fishermen	9.091

Table 8 shows the recommendations of social assistance programs for families in Yogyakarta city in general, including Community-based sanitation, Drinking water supply system, Social rehabilitation of uninhabitable houses, Pre-Employment card program, Cash assistance, and Family Hope Program. The recommendation for social assistance for the sample family in Yogyakarta city in Table 9 is the Pre-Employment Card Program that is shown in Table 10.

Table 9. Sample of Policy Mapping Based on Contributions of Features at Regional Levels in Yogyakarta City.

Dimension	Program	Percentage of Match
Housing	Community-based sanitation	100.0
Housing	Drinking water supply system	100.0
Housing	Social rehabilitation of uninhabitable houses	33.333
Economic	Pre-Employment card program	22.222
Social	Pre-Employment card program	22.222
Social	Cash assistance	10.0
Economic	Cash assistance	10.0
Health	Cash assistance	10.0
Health	Family Hope Program	6.25
Social	Family Hope Program	6.25
Education	Family Hope Program	6.25

Table 10. Sample of Policy Mapping Based on Contributions of Features at Family in Yogyakarta City.

Dimension	Program	Percentage of Match
Social	Pre-Employment Card	22.222
Economy	Pre-Employment Card	22.222

6. Discussion

Based on the Decree of the Coordinating Minister for Human Development and Culture of the Republic of Indonesia Number 32 Year 2022 concerning the General Guidelines for the Implementation of the Accelerated Program for the Eradication of Extreme Poverty, there are three main strategies for alleviating extreme poverty: (i) reducing the burden of community expenditures, (ii) increasing community income, and (iii) decreasing the number of poverty area. The results of modeling using SHAP indicate several majority characteristics that influence the welfare status of each family. Mapping between family characteristics to appropriate programs is not only at the family level but also at the regional level. This research will identify influencing factors in two districts and propose effective strategies for overcoming poverty through better and more accurate beneficiary targeting. The mapping of modeling results to government policy encompasses several programs that can be seen in the following table.

Table 11. List of Recommendation Programs to Accelerate the Eradication of Extreme Poverty at Regional Level

Region	*Strategy*	*Program*	*Mapping to Top 20 Features*	*Percentage of Match*
Kulonprogo Regency	Decreasing the number of poverty area	Social rehabilitation of uninhabitable houses	h_tfloor	33.333332
	Reducing the burden of community expenditures	Subsidy for 3-kilogram LPG	h_cookingfuel	14.285714
	Reducing the burden of community expenditures	Family Hope Program	h_nage04	6.25
Yogyakarta City	Decreasing the number of poverty area	Community-based sanitation	h_toiletype h_septic	100
	Decreasing the number of poverty area	Drinking water supply system	h_dwater	100
	Decreasing the number of poverty area	Social rehabilitation of uninhabitable houses	h_tfloor	33.333332
	Increasing community income	Pre-Employment card program	h_nage2064 h_nage0519	22.22221
	Reducing the burden of community expenditures	Cash assistance	h_njamkerja	10
	Reducing the burden of community expenditures	Family Hope Program	h_n_pendidikan_sma	6.25

There are similarities in strategies that are prioritized in improving family welfare based on feature contributions as shown in Table 11. For example, the dimension that has a large contribution in Kulonprogo Regency and Yogyakarta City is housing which is indicated by the floor area of the building residence per capita variable. This is also one of the criteria for the social rehabilitation program for uninhabitable houses. In addition, variables, such as final disposal place, types of toilets, and the main source of drinking water also have a high contribution to family welfare in Yogyakarta City. Therefore, program recommendations on the housing dimension are not only limited to social rehabilitation programs for uninhabitable houses but also community-based sanitation and drinking water supply systems.

Yogyakarta city has contribution features that lead to a strategy for increasing community income through a pre-employment card program. Features that have high contribution are the number of family members aged 20-64 years and aged 5-19 years. Based on the criteria for the pre-employment card program, the criterias are not only related to the age of the family member but also to working status. Meanwhile, in modeling results, the number of working-age and unemployed family members is not in the top 20 features. This finding can be used for improving the model through data evaluation so contribution features can be more aligned with the official criteria for each program.

The last strategy that is recommended for both regions is reducing the burden of community expenditure. This strategy is conducted through social assistance programs, social security, subsidies, price stability programs, and/or other programs that can reduce the burden of community expenditure. The modeling results show that comparison between features contribution and the program's criteria has a high percentage of matches in family hope programs for both regions, subsidy for 3-kilogram LPG for Kulonprogo Regency, and cash assistance for Yogyakarta city. The contribution features for family hope programs in Yogyakarta City are more related to education, while in Kulonprogo Regency are related to the social dimension. Fuel/main energy for cooking is a variable related to housing dimensions that contribute to the recommendation subsidy for 3-kilogram LPG in Kulonprogo Regency. The number of working hours of all working family members in Yogyakarta City has a significant contribution to the emergence of recommendations for direct cash assistance programs.

Overall, there are similarities and differences between program recommendations in the two districts that become study areas regarding the features or characteristics of a family that has a high contribution. By knowing the contribution features, policymakers can determine appropriate programs that can be conducted to targeted families in that region. The recommendations at the regional level will help to identify the most of the programs that are needed in that region.

7. Conclusion

Poverty is a multidimensional problem. Treatments and policies that are taken will differ depending on the characteristics of each region and family. The choice to intervene in a dimension to overcome the problem of poverty must be effective. The SHAP value gives the characteristics of the families in a region in general, more than that even the characteristics of the family individually. Based on the characteristics we can provide appropriate treatment or policies to overcome the problem and choose the right dimension to intervene. The social assistance programs that are provided to families will be more effective and impactful if they match the characteristics of the family.

By comparing two districts in Yogyakarta that represent high and low poverty, we get that program recommendations in the two districts may have both similarities and differences, depending on the characteristics of families with high contributions. By understanding these characteristics, policymakers can better determine the most appropriate programs to implement for target families in each area.

Acknowledgement

The authors would like to express their gratitude to BPS Statistics Indonesia for their full support in this research, for granting permission to use Regsosek data, and for allowing the publication of the study's results as scientific papers.

Appendix

Table 12. List of Variable (Processed) from Early Social Economic Registration (Regsosek) Data

id	Variable	Description	Dimention
0	h_house	The status of ownership of the residential building occupied	Housing
1	h_pcfloor	The floor area of the building residence per capita	Housing
2	h_tfloor	The widest type of floor	Housing
3	h_twall	bThe widest type of wall	Housing
4	h_troof	The widest type of roof	Housing
5	h_dwater	Main drinking water source	Housing
6	h_lighting	Main lighting source	Housing
7	h_cookingfuel	Fuel/main energy for cooking	Housing
8	h_toiltype	Type of toilet, ownership and use of defecating facilities	Housing
9	h_septic	Final Disposal Place	Housing
10	h_akses_internet	Families have internet access	Economy
11	h_nmale	The number of family members of the male sex	Social
12	h_nfemale	Number of Family Members of Female	Social
13	h_nage04	The number of family members aged 0-4 years	Social
14	h_nage0519	The number of family members aged 5-19 years	Social
15	h_nage2064	The number of family members aged 20-64 years	Social
16	h_nage65up	The number of family members aged 65 years and over	Social

id	Variable	Description	Dimention
17	h_hhcount	Number of family members	Social
18	h_ngrad_sd	The proportion of family members who completed elementary/equivalent	Education
19	h_ngrad_smp	The proportion of family members who completed junior high school/equivalent	Education
20	h_ngrad_sma	The proportion of family members who completed high school/equivalent	Education
21	h_ngrad_d	The proportion of family members who completed D1/D2/D3	Education
22	h_ngrad_s	Proportion of Family Members who completed D4/S1/S2/S3/Professional	Education
23	h_sec1_stat1	The number of family members who work in the agricultural sector & their own status	Economy
24	h_sec1_stat2	The number of family members who work in the agricultural sector & their status is assisted by non -permanent/unpaid workers	Economy
25	h_sec1_stat3	The number of family members who work in the agricultural sector & their status is assisted by permanent laborers/laborers paid	Economy
26	h_sec1_stat4	The number of family members who work in the agricultural sector & the status of labor/employees/employees	Economy
27	h_sec1_stat5	The number of family members who work in the agricultural sector & the status of free workers	Economy
28	h_sec1_stat6	The number of family members who work in the agricultural sector & the status of family workers/not paid	Economy
29	h_sec2_stat1	The number of family members who work in the industrial sector & their own status	Economy
30	h_sec2_stat2	The number of family members who work in the industrial sector & their status is assisted by non -permanent/unpaid workers	Economy
31	h_sec2_stat3	The number of family members who work in the industrial sector & their status is assisted by permanent laborers/laborers paid	Economy
32	h_sec2_stat4	The number of family members who work in the industrial sector & the status of labor/employees/employees	Economy
33	h_sec2_stat5	The number of family members who work in the industrial sector & the status of free workers	Economy
34	h_sec2_stat6	Number of family members who work in the industrial sector & their status of family workers/not paid	Economy
35	h_sec3_stat1	The number of family members who work in the service sector & status try alone	Economy
36	h_sec3_stat2	The number of family members who work in the service sector & status is trying to be assisted by non -permanent/unpaid workers	Economy
37	h_sec3_stat3	The number of family members who work in the service sector & status is trying to be assisted by permanent laborers/laborers paid	Economy
38	h_sec3_stat4	The number of family members who work in the service sector & the status of labor/employees/employees	Economy
39	h_sec3_stat5	The number of family members who work in the service sector & status of free workers	Economy

id	Variable	Description	Dimention
40	h_sec3_stat6	The number of family members who work in the service sector & status of family workers/not paid	Economy
41	h_ngangguan_penglihatan	The number of family members who have vision disorders	Health
42	h_ngangguan_pendengaran	The number of family members who have hearing loss	Health
43	h_ngangguan_belajar	The number of family members who have learning disorders	Health
44	h_ngangguan_gerak	The number of family members who have moving disturbances	Health
45	h_ngangguan_ingatan	The number of family members who have memory disorders	Health
46	h_ngangguan_perilaku	The number of family members who have behavioral disorders	Health
47	h_ngangguan_berbicara	The number of family members who have speaking disorders	Health
48	h_ngangguan_mengurus	The number of family members who have disturbances in taking care of themselves	Health
49	h_ngangguan_konsentrasi	The number of family members who have concentration disorders	Health
50	h_ngangguan_depresi	The number of family members who have depression disorders	Health
51	h_npenyakitkronis	The number of family members who have chronic diseases	Health
52	h_ngizi	Child nutritional condition	Health
53	h_njamkerja	The number of working hours of all working family members	Economy
54	h_nkerja	Number of Family Members Working	Economy
55	h_npunya_usaha	The number of family members who have their own business	Economy
56	h_nusaha	The number of own owned business oelh family	Economy
57	h_npekerja_dibayar	The number of workers is paid in a family -owned business	Economy
58	h_npekerja_tidak_dibayar	The number of workers is not paid in a family -owned business	Economy
59	h_nkepemilikan_izin	Business license ownership	Economy
60	h_nomzet_usaha_ultramikro	Number of businesses with ultramicro turnover	Economy
61	h_nomzet_usaha_mikro	The number of businesses with micro turnover	Economy
62	h_nomzet_usaha_kecil	The number of businesses with a small number of turnover	Economy
63	h_nomzet_usaha_menengah	The number of businesses with medium turnover	Economy
64	h_nomzet_usaha_besar	The number of businesses with a large turnover	Economy
65	h_kk_pendidikan_grad_sd	The head of the family is an elementary school graduate	Economy
66	h_kk_pendidikan_grad_smp	The head of the family is a junior high school graduate	Education
67	h_kk_pendidikan_grad_sma	The head of the family is a high school graduate	Education
68	h_kk_pendidikan_grad_d	The head of the family is a diploma graduate	Education
69	h_kk_pendidikan_grad_s	The head of the family is a graduate graduate	Education
70	h_npendidikan_sd	The number of family members with elementary school education	Education
71	h_npendidikan_smp	The number of family members with junior high school education	Education
72	h_npendidikan_sma	The number of family members with high school education	Education
73	h_npendidikan_d	The number of family members who have a diploma education	Education
74	h_npendidikan_s	The number of family members who are educated undergraduate	Education
75	h_nmale_age04	The number of male family members aged 0-4 years	Social
76	h_nmale_age0519	The number of male family members aged 5-19 years	Social
77	h_nmale_age2064	The number of male family members aged 20-64 years	Social
78	h_nmale_age65up	The number of male family members aged 65 years and over	Social
79	h_nfemale_age04	Number of female family members aged 0-4 years	Social
80	h_nfemale_age0519	The number of female family members aged 5-19 years	Social
81	h_nfemale_age2064	Number of female family members aged 20-64 years	Social

id	Variable	Description	Dimention
82	h_nfemale_age65up	The number of female family members aged 65 years and over	Social
83	h_nmale_age0519_kerja	The number of male family members aged 5-19 years	Economy
84	h_nmale_age2064_kerja	The number of male family members aged 20-64 years	Economy
85	h_nmale_age65up_kerja	The number of male family members aged 65 years and over	Economy
86	h_nfemale_age0519_kerja	The number of female family members aged 5-19 years	Economy
87	h_nfemale_age2064_kerja	The number of female family members aged 20-64 years	Economy
88	h_nfemale_age65up_kerja	The number of female family members aged 65 years and over	Economy
89	h_nage0519_kerja	The number of family members aged 5-19 years	Economy
90	h_nage2064_kerja	The number of family members aged 20-64 years	Economy
91	h_nage65up_kerja	The number of family members aged 65 years and over	Economy

References

[1] United Nations n.d. Transforming our world: the 2030 agenda for sustainable development [Internet] Department of Economic and Social Affairs Sustainable Development Goals [cited August 13, 2023] from https://sdgs.un.org/2030agenda

[2] Bappenas n.d. *tujuan-1*. SDGs [Internet] Kementerian PPN/Bappenas [cited August 13, 2023] Available from https://sdgs.bappenas.go.id/tujuan-1/

[3] Badan Pusat Statistik n.d. Kemiskinan dan ketimpangan [Internet] Badan Pusat Statistik [cited August 13, 2023] Available from https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html

[4] Lubi H 2017 Mengentaskan kemiskinan: Multidimensional Approach *Hermeneutika : Jurnal Hermeneutika* **3**(1)

[5] Huraerah A 2013 Strategi kebijakan penanggulangan kemiskinan di Indonesia *Jurnal Ilmu Kesejahteraan Sosial* **12**(1) pp 1-12 DOI: http://dx.doi.org/10.30870/hermeneutika.v3i1.2901

[6] SMERU 2021 July 26 Situasi kemiskinan Selama Pandemi [Internet] SMERU [cited August 13, 2023] Available from: https://smeru.or.id/id/article-id/situasi-kemiskinan-selama-pandemi

[7] Mufida N 2021 Efektivitas bantuan sosial tunai di kelurahan Purwosari kecamatan Purwosari kabupaten Pasuruan *Jurnal Sosial dan Sains* **1**(2) pp 82-92

[8] Noerkaisar N 2021 Efektivitas penyaluran bantuan sosial pemerintah untuk mengatasi dampak covid-19 di Indonesia *Jurnal Manajemen Perbendaharaan* **2**(1) pp 83-104 DOI: https://doi.org/10.59188/jurnalsosains.v1i2.23

[9] Aldino Putra A A 2018 Efektivitas pelaksanaan program bantuan sosial pada masyarakat di kota Palu *Jurnal elektronik Program Pascasarjana Universitas Tadulako* **6**(8)

[10] Badan Pusat Statistik 2022 Regsosek 2022 [Internet] Badan Pusat Statistik [cited August 27, 2023] Available from https://www.bps.go.id/regsosek/

[11] Alatas V, Banerjee A, Hanna R, Olken B A, and Tobias J 2012 June Targeting the poor: evidence from a field experiment in Indonesia *American Economic Review* **102**(4) pp 1206-1240 DOI: 10.1257/aer.102.4.1206

[12] Malta Z K and Sutikno 2019 Analisis karakteristik tingkat kesejahteraan di kota Surabaya menggunakan metode pohon klasifikasi *Jurnal Sains dan Seni ITS* **8**(2)

[13] Grosh M E and Baker J L 2013 Proxy means tests for targeting social programs *World Bank Book* DOI: https://doi.org/10.1596/0-8213-3313-5

[14] Taufiq N and Mariyah S 2021 Pendekatan model machine learning dalam pemeringkatan Status Sosial Ekonomi Rumah Tangga di Indonesia *Prosiding Seminar Nasional Official Statistics 2021* **2021(1)** pp 1044-1053 DOI: https://doi.org/10.34123/semnasoffstat.v2021i1.1018

[15] Poulin C, Trimmer J, Press-Williams J, Yachori B, Khush R, Peletz R, and Delaire C 2022 Performance of a novel machine learning-based proxy means test in comparison to other methods for targeting pro-poor water subsidies in Ghana *Development Engineering* **7** DOI: https://doi.org/10.1016/j.deveng.2022.100098

[16] Aiken E, Bellue S, Karlan D, Udry C, and Blumenstock J E 2022 Machine learning and phone data can improve targeting of humanitarian aid *Nature* **603** pp 864-870 DOI: https://doi.org/10.1038/s41586-022-04484-9

[17] Kambuya P 2020 Better model selection for poverty targeting through machine learning: A Case Study in Thailand *Thailand and The World Economy* **38**(1) pp 91-116

[18] Hancock J T and Khishgiftaar T M 2020 CatBoost for big data: an interdisciplinary review. *Journal of Big Data* **7** p 94 DOI: https://doi.org/10.1186/s40537-020-00369-8

[19] Breiman L, Friedman J, Stone C J, and Olshen R A 1984 *Classification and Regression Trees* (Taylor & Francis) DOI: https://doi.org/10.1201/9781315139470

[20] Breiman L 2001 Random forests *Machine Learning* **45** pp 5-32 DOI: 10.1023/A:1010950718922

[21] Friedman J 2000 Greedy function approximation: a gradient boosting machine *The Annals of Statistics* **29** pp 1189-1232

[22] Prokhorenkova L, Gleb G, Vorobev A, Dorogush A V, and Gulin A 2018 CatBoost: unbiased boosting with categorical features *Proceedings of the 32nd International Conference on Neural Information Processing Systems* **31** pp 6639–6649

[23] Dorogush A V, Ershov V, and Gulin A 2018 CatBoost: gradient boosting with categorical features support DOI: 10.48550/arXiv.1810.11363

[24] Berrar D 2018 Cross-validation *Reference Module in Life Sciences* (Elsevier) DOI: 10.1016/B978-0-12-809633-8.20349-X

[25] Lundberg S M and Lee S-I 2017 A unified approach to interpreting model predictions *Proceedings of the 31st International Conference on Neural Information Processing Systems* **30** pp 4768–4777