Urban vs Rural: A Machine Learning Approach to Household Classification in the Philippines

By Rachel Joy Baldo

Abstract: The swift shift of the Philippines' population towards urban areas has resulted in notable variations in the economic activities of households in urban and rural settings. Grasping these variations is vital for policymakers, researchers, and social scientists as it can guide the formulation of effective strategies aimed at economic growth, poverty alleviation, and the efficient distribution of resources. This research applies machine learning to classify Philippine households as Urban or Rural based on their socioeconomic activities. Harnessing the 2021 Family Income and Expenditure Survey (FIES) dataset from the Philippine Statistics Authority (PSA), household income, expenditure, and consumption patterns were analyzed for valuable insights. Using Jupyter Python, the research trains and evaluates five binary classification models on the data and selects XGBoost as the best model—with an accuracy score of 0.67 and a ROC-AUC score of 0.72. SHAP values effectively visualized, identified, and interpreted the most important features influencing the model's predictions. From the analysis, Urban households tend to have higher Non-Food and Food Expenditures and higher Per Capita Income. Meanwhile, Rural households tend to have larger Family Sizes and higher Total Income due to more diverse income sources. Future research can benefit from exploring other machine learning techniques, incorporating spatial and temporal data, and validating the model with external data sources.

1. Introduction

1.1 Background

The ongoing urban transition in the Philippines has reshaped economic behaviors among households residing in urban and rural areas. As the urban population in the country grows—from 51.2% in 2015 to 54.0% in 2020 (PSA 2022)—understanding the nuanced patterns and trends of these behaviors becomes critical. By examining the main economic activities of Philippine households, the government can inform targeted strategies for suitable economic policies, income inequality reduction, and sustainable development (World Bank 2022; Monsura 2021).

Previous studies have shown differences between urban and rural lifestyle choices, particularly in necessities like food consumption (de Luna and Bullecer 2020). Additionally, there are significant urban-rural disparities in income sources, with policies often favoring major cities in the Philippines, such as Metro Manila (Chua et al. 2015). Utilizing the Family Income and Expenditure Survey (FIES) 2021 dataset, the study aims to classify Philippine households as urban or rural based on their income, expenditure, and consumption patterns using machine learning. The goal is to reveal a more fundamental analysis of the primary factors present in urban and rural communities.

1.2 Importance

One of the main causes of poverty in the Philippines is the uneven distribution of economic growth across the regions (Hara 2023). Despite overall economic growth, poverty still persists because the gains are not reaching those who need it most (Albert et al. 2017). Public policies rely on precise data and valuable insights. By analyzing urban and rural households' economic activities, policymakers can gain the information needed to create holistic interventions. This will guide efficient resource allocation to meet the needs of both urban and rural populations (Andriesse 2017). Understanding the economic patterns among urban and rural will help balance urban growth with rural conservation. Bridging the urban-rural divide will ensure equal economic access to opportunities and resources for all Filipinos.

The issue of disparities between urban and rural areas is not unique to the Philippines, but is a global concern. Insights gained from this research can inform policy development on an international scale. Comparative studies are instrumental in understanding shared challenges and devising innovative solutions (Avoyan 2022). Addressing these disparities is also pivotal for achieving the United Nations Sustainable Development Goals—contributing to objectives such

as No Poverty, Reduced Inequalities, and Sustainable Cities and Communities (United Nations 2015).

Incorporating machine learning methods into the research introduces new possibilities. Machine learning models are adept at revealing hidden patterns within data, forecasting trends, and providing insights that surpass traditional analytical methods. Predictive models facilitate effective and efficient decision-making, such as forecasting future migration patterns, which is vital for urban and rural planning (Chakraborty et al. 2022)

1.3 Research Objective

This research aims to identify the factors that classify households in the Philippines as either urban or rural, based on their income, expenditure, and consumption habits. By leveraging the 2021 Family Income and Expenditure Survey dataset provided by the Philippine Statistics Authority and employing machine learning techniques to analyze these features, the study seeks to uncover underlying factors that influence these economic behaviors and how these factors contribute to the urban-rural divide. For this purpose, different binary classification models are trained and evaluated on the FIES Data to determine the best classifier for classifying urban-rural household.

2. Methods

2.1 The FIES Dataset

The data used in this study were derived from the Family Income and Expenditure Survey 2021 Volume 2 database, a nationwide survey conducted by the Philippine Statistics Authority (publicly available at https://psada.psa.gov.ph/catalog/FIES/about). It is designed to collect information on the income and expenditure of families in the Philippines. The FIES provided a comprehensive set of data that was essential for evaluating the living standards and economic conditions of Filipino households. The FIES has been conducted every three years since 1985, but starting in 2023, it will be conducted biennially. The 2021 FIES survey used a sample size of 165,029 households, considering the 17 regions of the country. Further stratification was performed using geographic groups such as provinces, highly urbanized cities, and independent component cities.

Included in the FIES dataset were thirty-seven variables but for this study, seven relevant variables were used for easier data manipulation and model testing:

Family Size (FSIZE) – The number of individuals in a household significantly influences consumption patterns and financial priorities. This variable is essential for analyzing how economic behaviors change with family size.

Per Capita Income (PCINC) –This variable measures the total income dived by the family size of a household. It is a key indicator of the economic status of a family and is used to compare the wealth of different populations.

Total Expenditure (TTOTEX) – The sum of total food expenditure and total non-food expenditure, this aggregate of all expenditures gives a comprehensive view of a household's financial outflows, including both essential and discretionary spending.

Total Food Expenditure (TFOOD) – This variable captures the total amount spent by a household on food. It reflects the cost of living, dietary preferences, and an indicator of food security and nutrition levels.

Total Income (TOINC) – This is the sum of all income sources for a household, including wages, business profits, and remittances. It is a direct measure of financial resources available to a household.

Total Non-Food Expenditure (TNFOOD) – This variable includes all other expenses outside of food, such as housing, utilities, education, and healthcare. It provides insight into the allocation of resources to different life necessities

Urban or Rural (URB) – Classifies households based on their urban or rural location, affecting lifestyle and economic opportunities.

2.2 Data Preprocessing and Machine Learning Modeling

After loading the dataset into the Jupyter Python notebook, an exploratory analysis was conducted to check the data shape and data types, handle missing row values, and drop feature columns that were not relevant to this study. Descriptive analysis provided the mean, median, standard deviation values of the datasets. These values showed the distribution, variability, and central tendency of the dataset.

The remaining dataset was visualized to easily determine the dispersion of outliers within each feature. The dataset was further examined for class imbalance to ascertain whether the classes required oversampling or undersampling. Since the 'URB' column had legends 1 for Urban and 2 for Rural, the researcher replaced the legends with 0 representing Urban households and 1 representing Rural households, in accordance with the binary classification nature of this study. Of the 165,029 total, 79,202 were classified as Urban and 85,827 as Rural, which did not indicate an imbalance. A normalization per variable using MinMaxScaler was performed so that each variable range is from 0 to 1. Then the data were split into 80% training set and 20% test set using stratified sampling, so both urban and rural households were accordingly proportioned.

The objective of this study was to explore the influence of key features—'FSIZE', 'PCINC', 'TTOTEX', 'TFOOD', 'TOINC', and 'TNFOOD'—on urban-rural 'URB' economic behaviors and to assess how these features contribute to balancing urban growth with rural conservation. Machine learning (ML) algorithms are commonly used to discover the complex and inherent connections between input and output data. A thorough 10-fold cross validation, grid search, and hyperparameter tuning process were undertaken for various ML models, including Logistic Regression, K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree, and XGBoost. To identify the most effective model for the classification task, metrics like Precision, Recall, F1-Score, and Accuracy were evaluated. Precision is the ratio of correctly predicted positive observations to the total predicted positives. Recall is the ratio of correctly predicted positive observations to all observations in actual class. F1-score measures the balance between precision and recall. Accuracy measures the overall correctness of the model's predictions.

3. Results

3.1 Machine Learning Modeling Results

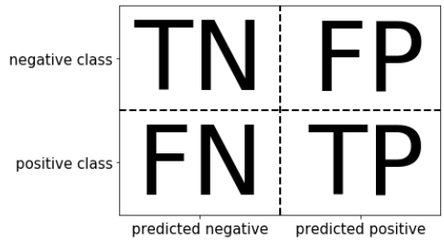
From the classification metrics, the best classifier for the task were Random Forest (RF) and XGBoost in terms of accuracy, with 0.67 accuracy score. Table 1 summarizes the metric results.

Table 1. Summary of the best validation performance per ML model after 10-fold cross validation and grid search. **Bold text** indicates the best ML model performance based on accuracy.

Class	ML Model	Validation Performance				Best Parameters
		Precision	Recall	F1-Score	Accuracy	
0 - Urban	Logistic Regression	0.68	0.47	0.56	0.64	C': 100, 'penalty': 'l2', 'solver': 'liblinear'
	KNN	0.63	0.62	0.62	0.64	'n_neighbors': 11
	RF	0.65	0.66	0.66	**0.67**	n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_depth': 10
	Decision Tree	0.64	0.66	0.65	0.66	max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 5
	XGBoost	0.66	0.66	0.66	**0.67**	learning_rate': 0.01, 'max_depth': 7, 'n_estimators': 500
1 - Rural	Logistic Regression	0.62	0.80	0.70	0.64	C': 100, 'penalty': 'l2', 'solver': 'liblinear'

KNN	0.65	0.66	0.66	0.64	'n_neighbors': 11
RF	0.68	0.67	0.68	**0.67**	n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_depth': 10
Decision Tree	0.68	0.66	0.67	0.66	max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 5
XGBoost	0.68	0.68	0.68	**0.67**	learning_rate': 0.01, 'max_depth': 7, 'n_estimators': 500

A Confusion Matrix was then applied to further evaluate the performance of each ML model. This matrix is a two-by-two table that records the counts of true positives, true negatives, false positives, and false negatives. This matrix goes beyond mere accuracy to provide a more detailed analysis of a model's performance, highlighting the types of errors made and offering insights into precision, recall, and other metrics.



(a)

(b)

Figure 1. **(a)** Interpretation of confusion matrix, showing the placement for True Negative, False Positive, False Negative, True Positive. **(b)** Confusion matrix for each classification model.

The Confusion Matrix above revealed XGBoost classifier performed well in predicting 10,362 True Negative (0) and 11,655 True Positives (1). Overall, XGBoost outperformed the other ML models in terms of performance consistency.

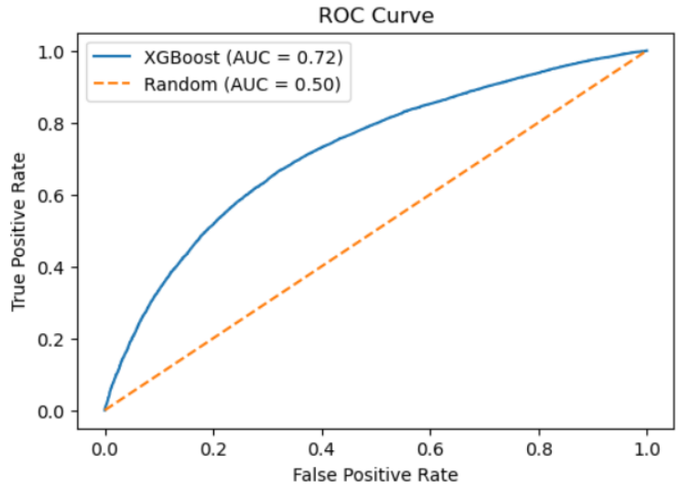


Figure 2. Area Under the Receiver Operating Characteristic **(**ROC-AUC) Curve for XGBoost model.

Following the analysis with the Confusion Matrix, which highlighted XGBoost as the superior ML classifier, a ROC Curve was constructed for XGBoost. This graphical representation serves to visualize the classifier's performance across various threshold settings. The higher the ROC-AUC value, the better the model performance in signifying that the model has a good measure of separability, and of distinguishing between the classes—Urban (0, Negative) and Rural (1, Positive)—with greater clarity.

3.2 Feature Importance Analysis

For post-hoc analysis and to identify the level of importance with which features affect the classification of urban-rural households, the Shapley Additive Explanations (or SHAP values) was applied. These values distribute the prediction outcome among the features fairly (Lundberg and Lee, 2017). SHAP values represented the average impact of a feature across every possible combination of features, providing a global view of feature importance. This resulted in a hierarchy of features—ranked by their importance, with the most significant features at the top—indicating their influence on the XGBoost's predictions.

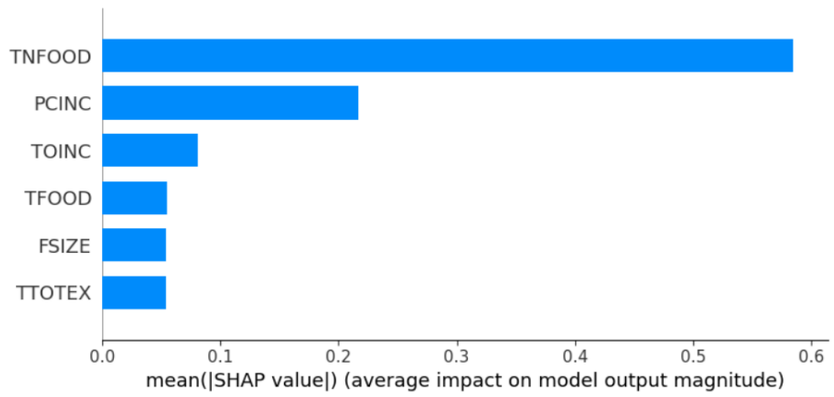


Figure 3. Bar graph of mean SHAP values for the features in XGBoost classifier.

The features TNFOOD, PCINC, TOINC, TFOOD, FSIZE, and TTOTEX were listed on the y- axis, representing the feature variables in the model. The x-axis shows the mean SHAP value, which quantifies the average impact of each feature on the model's output magnitude. The length of each bar represents the importance of the corresponding feature, with longer bars indicating greater importance. In this graph, the feature TNFOOD had the longest bar, suggesting it had the highest mean SHAP value and, therefore, had the most significant impact on the model's predictions. The other features followed in descending order of importance.

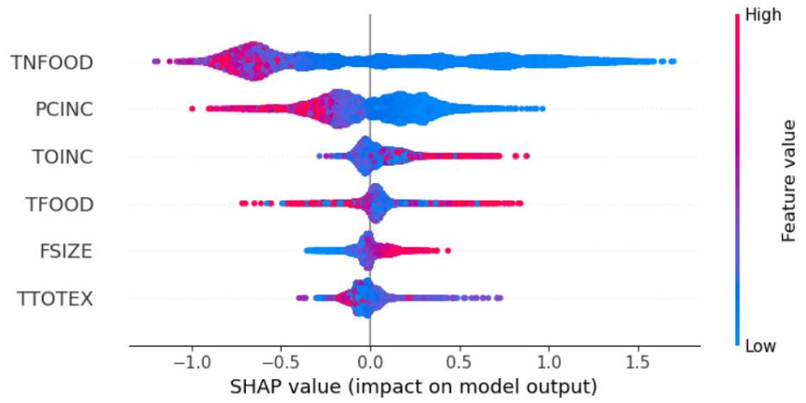


Figure 4. Summary density scatter plot of SHAP values for each feature to identify how much impact each feature has on the model output for individuals in the validation dataset.

The density scatter plot showed the distribution of SHAP values for each feature on the x-axis, with the y-axis representing the individual features. The density of the scattered dots indicated how often a particular SHAP value occurs. The color of the dots represented the feature value, with blue representing lower values and red representing higher values. Whether each feature had a positive or negative impact on the model output depends on its position relative to the zero line on the horizontal axis. Additionally, features with positive SHAP values positively impact the prediction, while those with negative values have a negative impact. The magnitude of the SHAP value is a measure of how strong the effect is.

Figure 5. SHAP Force plot to interpret how the XGBoost model arrived at the prediction.

The objective of this study was to determine whether individuals resided in Urban (0) or Rural (1) households based on these features. The SHAP force plot above illustrated that the XGBoost model assigned a score of 0.36. A higher score tends to predict '1', while a lower score leans towards '0'. Since the model score was lower, its predictions lean towards '0' which was the Urban class in the dataset. The features that significantly influenced the prediction were depicted in red and blue. Red indicated features that increased the model's score, while blue signified features that decreased it. Features with a greater impact on the score were positioned nearer to the boundary separating red and blue. The magnitude of their influence was represented by the length of the bar (Steele 2022).

4. Discussions

The aim of this research was to leverage machine learning methods to pinpoint important variables that determine whether households in the Philippines were classified as urban or rural, based on their income, spending, and consumption patterns. The intention was to conduct a deeper examination of the predominant elements found in urban and rural settings and to unravel the fundamental factors that drive these economic behaviors. The 2021 FIES Dataset was analyzed to shed light on the factors that contribute to the urban-rural disparity. The XGBoost classifier outperformed the other ML models after hyperparameter tuning and metric evaluation; Shapley Additive Explanations (SHAP) analysis visualized the average impact of each feature on the model's output magnitude.

The TNFOOD (Total Non-Food Expenditure) feature represented spending on non-food items. This means that, on average, changes in non-food expenditure have a substantial effect on the model's output. From the SHAP values, TNFOOD had many positive blue dots, suggesting higher values of TNFOOD tend to decrease the model's prediction score. This means that higher TNFOOD values were pushing the model's prediction towards '0', thus higher non-food

expenditures were characteristic of Urban households. This insight aligns with the general understanding that urban households tend to have higher living costs and more diverse needs beyond basic food items, leading to increased spending on non-food goods and services. Previous studies have noted factors like the rise in expenditure on durables, education, healthcare and other services (Basole and Basu 2015), and higher market shocks (Valera et al. 2022). Additionally, urban areas typically offer more employment opportunities and higher incomes (Connolly et al. 2022), which can result in greater disposable income and the capacity for increased spending on non-food items.

Interestingly, PCINC (Per Capita Income), which measured the total income divided by family size, was also associated with urban households. More positive blue dots on the SHAP density scatter plot indicated that higher values of PCINC lower the model's prediction score, thus pushing the model's prediction towards '0'. This result was consistent with studies showing that higher per capita income was a characteristic of urban households (Mi et al. 2020), and that urban areas typically had higher income levels due to greater access to jobs, higher wages, and more diverse economic opportunities (Nguyen et al. 2015) compared to rural areas.

For TOINC (Total Income) feature, or the sum of all income sources for a household, including wages, business profits, and remittances, most dots were red and on the positive side of SHAP plot. It suggested that higher values of TOINC tend to increase the model's prediction score. In other words, higher TOINC values were pushing the model's prediction towards '1'. Some studies have found that rural areas had more diverse and resilient sources of income than urban areas, especially during the COVID-19 pandemic. One study discovered that rural communities had more access to natural resources and social capital, which enhanced their livelihood resilience and enabled them to adopt different livelihood strategies (Liu 2020). Another study had found that rural households relied on a mix of farm, nonfarm income, and remittances to cope with shocks and sustain their livelihoods (Quimba and Estudillo 2018).

TFOOD (Total Food Expenditure) was the variable that captures the total amount spent by a Filipino household on food. It reflected the cost of living, dietary preferences, and food security indicators and nutrition levels. The feature had red dots on both negative and positive sides, suggesting TFOOD values can both increase and decrease the model's prediction score. But in Figure 5 for SHAP force plot, TFOOD was interpreted as having higher impact on the model's prediction towards '0'. Some studies found out that urban households spend more on food than rural households and that urban food basket composition was more diverse (Bairagi et al. 2022). A 2023 quarterly study conducted by PSA also highlighted urban households have higher expenditures on restaurants and hotels.

FSIZE (Family Size) represented the number of individuals in a household, this was essential for analyzing how economic behaviors change with family size. On the SHAP density scatter plot, FSIZE had almost equal number of positive red dots and negative blue dots, showing that this feature had a complex relationship with the model's prediction. SHAP force plot provided a clear interpretation of this feature, where positive red dots outweighed the negative blue dots. So FSIZE values had a positive impact on the prediction, pushing it towards '1'—where higher FSIZE values were associated with Rural households. One study analyzed the household family size and structure of the Philippines and found that rural households had more children and extended family members (Monsura 2021).

TTOTEX (Total Expenditure) or the sum of total food expenditure and total non-food expenditure, also showed a more complex relationship with the model's prediction. The scattered red and blue dots on the positive side indicated both higher and lower TTOTEX values have a positive impact on the prediction, pushing it towards '1'. But this feature had a very small SHAP value based on the force plot result, thus the plot showed TTOTEX at the very end of the red bar.

5. Conclusion

This research aims to develop and evaluate a machine learning model that can classify households in the Philippines as 0=Urban or 1=Rural based on their socioeconomic characteristics. Using the 2021 Family Income and Expenditure Survey (FIES) dataset, the study applied XGBoost to train and test the data. Results showed that the XGBoost model achieved moderately high

accuracy, precision, recall, and f1-score, indicating its effectiveness in distinguishing urban and rural households. For post-hoc analysis, utilizing SHAP values effectively visualized, identified, and interpreted the most important features influencing the model's predictions. Total Non-Food Expenditure (TNFOOD), Per Capita Income (PCINC), and Total Income (TOINC) were the top three features with the most impact and contribution to the model's output. These features reflect the differences in consumption patterns, living standards, and demographic factors between Urban and Rural households. With this, suggestions for further research, such as exploring other machine learning techniques, incorporating spatial and temporal data, and validating the model with external data sources are beneficial and advantageous.

By understanding the granular-level factors influencing socioeconomic behaviors, policymakers, researchers, and social scientists can design more holistic programs, laws, and interventions to benefit both the grassroots of Urban and Rural communities. To achieve sustainable development, policies should take into account the environmental and natural resource implications of urbanization and rural preservation. To improve food security and nutrition, policies should address the accessibility and affordability of food products in both urban and rural settings. Only six features were deemed necessary for this classification research, which may not capture the full complexity of the variables. Other aspects, such as geographic location, access to better infrastructures and services, social and cultural norms, and environmental factors should also be considered. Alas, the computing power of the researcher's laptop cannot handle bigger data.

References

1. Albert, Jose Ramon G.; Dumagan, Jesus C.; Martinez, Arturo Jr. (2015) : Inequalities in Income, Labor, and Education: The Challenge of Inclusive Growth, PIDS Discussion Paper Series, No. 2015-01, Philippine Institute for Development Studies (PIDS), Makati City
2. Andriesse, E. (2017). Regional disparities in the Philippines: structural drivers and policy considerations. *Erdkunde*, *71*(2), 97–110. https://doi.org/10.3112/erdkunde.2017.02.01
3. Avoyan, E. (2022). Collaborative Governance for Innovative Environmental Solutions: Qualitative Comparative Analysis of Cases from Around the World. *Environmental Management*, *71*(3), 670–684. https://doi.org/10.1007/s00267-022-01642-7
4. Bairagi, S., Zereyesus, Y., Baruah, S., & Mohanty, S. (2022). Structural shifts in food basket composition of rural and urban Philippines: Implications for the food supply system. *PLOS ONE*, *17*(3), e0264079. https://doi.org/10.1371/journal.pone.0264079
5. Basole, A., & Basu, D. (2015). Non-Food Expenditures and Consumption Inequality in India. *Economic and Political Weekly*, *50*(36), 43–53. https://blogs.umb.edu/amitbasole/files/2013/10/Basole_Basu_NonFood_Expenditures_and_Consumption_Inequality_in_India-1qjei9y.pdf
6. Chakraborty, A., Sikder, S. K., Omrani, H., & Teller, J. (2022). Cellular Aut006Fmata in Modeling and Predicting Urban Densification: Revisiting the Literature since 1971. *Land*, *11*(7), 1113. https://doi.org/10.3390/land11071113
7. Chua, K. K. T., Limkin, L., Nye, J. V. C., & Williamson, J. G. (2015). Urban-rural income and wage gaps in the Philippines: measurement error, unequal endowments, or factor market failure? *He Philippine Review of Economics Vol. LII No. 2, December 2015 Pp. 1-21*, *52*(2), 1–21.
8. Connolly, M., Shan, Y., Bruckner, B., Li, R., & Hubacek, K. (2022). Urban and rural carbon footprints in developing countries. *Environmental Research Letters*, *17*(8), 084005. https://doi.org/10.1088/1748-9326/ac7c2a
9. De Luna, K. L. G. D., & Bullecer, E. R. (2020). Rural and urban differences in household food insecurity and Diet Diversity of preschool Children (PSC) in Occidental Mindoro. *Acta Medica Philippina*, *54*(5). https://doi.org/10.47895/amp.v54i5.2254
10. Hara, M. (2023). Educational reform for middle-income trap under digitalization: Culprits, challenges, and strategies in the Philippines. SocioEconomic Challenges, 7(3), 200-218. https://doi.org/10.61093/sec.7(3).200-218.2023.
11. *Household Final Consumption Expenditure, Q1 2020 to Q3 2023 Growth Rates | Philippine Statistics Authority (2023). Republic of the Philippines*. (n.d.). https://psa.gov.ph/statistics/national-accounts/sector/Household%20Final%20Consumption

12. Liu, W., Li, J., Ren, L., Xu, J., Li, C., & Li, S. (2020). Exploring livelihood resilience and its impact on livelihood strategy in rural China. *Social Indicators Research*, *150*(3), 977–998. https://doi.org/10.1007/s11205-020-02347-2
13. Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *Neural Information Processing Systems*, *30*, 4768–4777. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
14. Mi, Z., Zheng, J., Meng, J., Ou, J., Hubacek, K., Liu, Z., Coffman, D., Stern, N., Liang, S., & Wei, Y. (2020). Economic development and converging household carbon footprints in China. *Nature Sustainability*, *3*(7), 529–537. https://doi.org/10.1038/s41893-020-0504-y
15. Monsura, M. P. (2021). *Socioeconomic Characteristics of Households, Government Programs on Human Capital, and Natural Shocks as Determinants of Philippine Household Income Mobility*. Asia-Pacific Social Science Review - De La Salle University. https://www.dlsu.edu.ph/wp-content/uploads/pdf/research/journals/apssr/2021-December-vol21-4/1-prelim.pdf
16. Nguyen, L., Raabe, K., & Grote, U. (2015). Rural–Urban migration, household vulnerability, and welfare in Vietnam. *World Development*, *71*, 79–93. https://doi.org/10.1016/j.worlddev.2013.11.002
17. Quimba, F., & Estudillo, J. (2016). Sources of income in rural Philippines: the role of population pressure, urbanization and infrastructure development. *Asia Pacific Journal of Multidisciplinary Research*, *Vol. 6*(No.1), 37–45.
18. Steele, M. (2022, March 30). SHAP force plots for classification - MLearning.ai - medium. *Medium*. https://medium.com/mlearning-ai/shap-force-plots-for-classification-d30be430e195
19. *United Nations (2015). Transforming our World: The 2030 Agenda for Sustainable Development | Department of Economic and Social Affairs*. (n.d.). https://sdgs.un.org/publications/transforming-our-world-2030-agenda-sustainable-development-17981
20. *Urban Population of the Philippines (2020 Census of Population and Housing) | Philippine Statistics Authority | Republic of the Philippines*. (n.d.). https://psa.gov.ph/content/urban-population-philippines-2020-census-population-and-housing
21. Valera, H. G. A., Mayorga, J., Pede, V. O., & Mishra, A. K. (2022). Estimating food demand and the impact of market shocks on food expenditures: The case for the Philippines and missing price data. *Q Open*, *2*(2). https://doi.org/10.1093/qopen/qoac030
22. World Bank (2022). *Overcoming poverty and inequality in the Philippines: Past, Present, and Prospects for the Future*.