RJOAS: Russian Journal of Agricultural and Socio-Economic Sciences
ISSN 2226-1184 (Online) | Issue 7(151), July 2024

UDC 332

PREDICTING BEEF CONSUMER CHARACTERISTICS IN INDONESIA USING A COMBINATION OF RESAMPLING AND MACHINE LEARNING TECHNIQUES

Hadi Sholih Nugroho[1,2], **Chung Rebecca H.**[1]*
[1]Department of Tropical Agriculture and International Cooperation,
National Pingtung University of Science and Technology, Pingtung, Taiwan
[2]Ministry of Agriculture, Jakarta 12540, Indonesia
*E-mail: rebecca@mail.npust.edu.tw

ABSTRACT
Although per capita beef consumption in Indonesia remains relatively low, the demand for beef has increased significantly over the past three decades. With domestic production unable to keep pace with this growing demand, the country has resorted to imports from other countries. This study aims to predict beef consumers and identify their characteristics in Indonesia using various algorithms combined with resampling techniques to handle an imbalanced dataset. A total of 33,361 observations were collected from the National Socioeconomic and Expenditure Survey conducted in 2021. Based on criteria of precision, sensitivity and F-score, oversampling produces higher accuracy than undersampling and no sampling. Among the six classifiers used to predict beef consumption in Indonesia, random forest and support vector machine outperformed others. Income level is identified as the most powerful factor for predicting beef consumption, followed by culture, house size, education level, age and consumption of poultry meat. These findings offer valuable insights for stakeholders involved in the beef industry, empowering them to devise more effective strategies. Furthermore, government can leverage these results to formulate macro policies pertaining to the beef industry, ensuring alignment with evolving consumer demand and fostering sustainable sectoral growth.

KEY WORDS
Beef consumption, F-score, oversampling, random forest, support vector machine.

Beef is one of the most important animal protein sources other than egg and chicken meat in Indonesia (Anindita et al., 2020). Beef consumption is mainly required by households, bakso (meatballs, a very popular food in Indonesia) sellers, food catering, the food industry, and tourism (Agus and Widi, 2018). Although per capita beef consumption of 0.009 kg per week is significantly lower than chicken consumption, which is 0.126 kg (Ministry of Agriculture of Indonesia, 2022), beef remains an indispensable in various well-known Indonesian cuisines such as bakso, rendang (caramelized beef curry) and gulai (beef goulash).

In Indonesia, beef prices exceed those in the international market (Hadi and Chung, 2022), and are considerably expensive than the alternative protein sources such as chicken and eggs (PIHPS, 2023). Consequently, beef consumption remains limited to specific groups of people routinely. For the majority of Indonesians, beef is served for special occasions, such as religious or wedding celebrations, seen as a symbol of exclusivity and luxury. Notably, households categorized as extremely impoverished, impoverished, almost impoverished, and socioeconomically disadvantaged, accounting for 40% of the population, find beef largely inaccessible (Khoiriyah et al., 2020). Nonetheless, the demand for beef in Indonesia continues to increase annually, prompting imports from other countries (Hadi and Chung, 2022; Hamdan et al., 2019; Hovhannisyan and Devadoss, 2020). Accordingly, this raises several questions: 1) Which households regularly consume beef in their diets? 2) What factors drive beef consumption in Indonesia?

Previously, there has been extensive discussion concerning the factors that influence beef consumption across various studies conducted in several countries. Using logit regression, Uzunöz and Karakaş (2014) determine socioeconomic factors influencing beef

consumption in Turkey. The key results revealed that income positively affected beef consumption, while education level and household size were negatively associated with beef consumption. Zhang et al. (2018) reported that the meat consumption pattern in China is subject to alteration due to fluctuations in income levels based on advanced linear approximated almost ideal demand system (AIDS) analysis. Yildirim and Ceylan (2008) used independent-sample t-students, one-way ANOVA, chi-square and linear regression statistical tests to elucidate the significant impact of income on consumption patterns in Turkey. In addition, their findings revealed that urban families placed greater emphasis on factors related to habits and nutritional content, whereas affordability emerged as the primary factor shaping the meat preferences among rural households. Furthermore, socio-demographic variables such as gender, age, income, and education level were identified as significant contributors to consumption (Popoola et al., 2022). Similarly, the statistical importance of respondents' educational level, employment, income, and age across varying levels of beef intake has been observed (Jabo and Zaharadden, 2018). Recently, Garaus and Garaus (2023) underscores the impact of alternative sources of protein on beef consumption in the U.S.

Machine learning algorithms have been successfully applied in many fields, including the health sector (Davagdorj et al., 2020; Kwekha-Rashid et al., 2023), medical practice (Abbasi and Goldenholz, 2019; Cuocolo et al., 2019), hotel hospitality (Sann et al., 2022), precision and sustainable agriculture (Sharma et al., 2020; Sharma et al., 2021), and chemistry (Townsend et al., 2020)(Townsend et al., 2020), among others. However, despite their widespread utilization, the application of machine learning techniques in predicting beef consumers remains largely unexplored and under-discussed. Therefore, this study aims to predict beef consumers in Indonesia using several machine learning techniques. Specifically, the objectives of this study encompass three key areas. First, it aims to delineate the patterns of beef consumption within the context of Indonesia. Then, the study seeks to conduct a comparative analysis of different machine learning techniques employed in predicting beef consumers among Indonesian population. Finally, it endeavors to delve into the factors influencing beef consumption and profile the characteristics of the beef consumers in Indonesia. Through these objectives, the study aims to offer insights into beef consumption pattern, contribute to the refinement of predictive models, and shed light on the determinants shaping consumer behaviors in the Indonesian beef market.

Understanding consumer characteristics empowers marketers to effectively target distinct population segments. By identifying the preferences, needs, and behaviors of diverse consumer groups, marketers can customize their strategies to resonate with each segment individually, ultimately leading to heightened conversion rates and sales. Furthermore, policymakers can leverage insights gleaned from forecasting consumer traits to formulate regulations and policies pertaining to the beef industry. This study is also expected to enrich the literature on the utilization of machine learning and resampling techniques in predicting consumption behavior.

MATERIALS AND METHODS OF RESEARCH

The experimental design for predicting beef consumer characteristics consists of three key stages. In the first stage, this study removes missing value, smooths data distribution and eliminates outliers. Subsequently, resampling techniques are applied to mitigate potential biases in classifier performance stemming from imbalanced datasets. Finally, multiple machine learning classification algorithms are employed to predict beef consumer characteristics. Various criteria are utilized to evaluate the accuracy of the classifier predictions to determine the most suitable classifier for the task. For model evaluation, the dataset was randomly split into 60% for the training set, 20% for the validation set, and 20% for testing (Wendler and Gröttrup, 2021a). The proposed flow of experimental designs is depicted in Figure 1.

Real-world datasets often contain valuable information but may lack the appropriate format requirement for data mining procedures. Data preprocessing on big data is a crucial

stage in the data analysis pipeline, as it prepares raw data for subsequent analysis and modelling. Big data, characterized by its large size and complexity, typically requires specialized techniques and tools for processing, often involving distributed computation to manage the volume and velocity of the data. The first step of data preprocessing in this study involves data cleaning. Given the size and diverse sources of big data, it is more susceptible to errors, missing values, and outliers. Data cleaning encompasses tasks such as handling missing values, rectifying mistakes, and addressing outliers. The second step focuses on data transformation, where the data is converted into a format suitable for analysis. Common transformations include normalization (scaling data to a standard range), log transformation, and one-hot encoding for categorical variables.

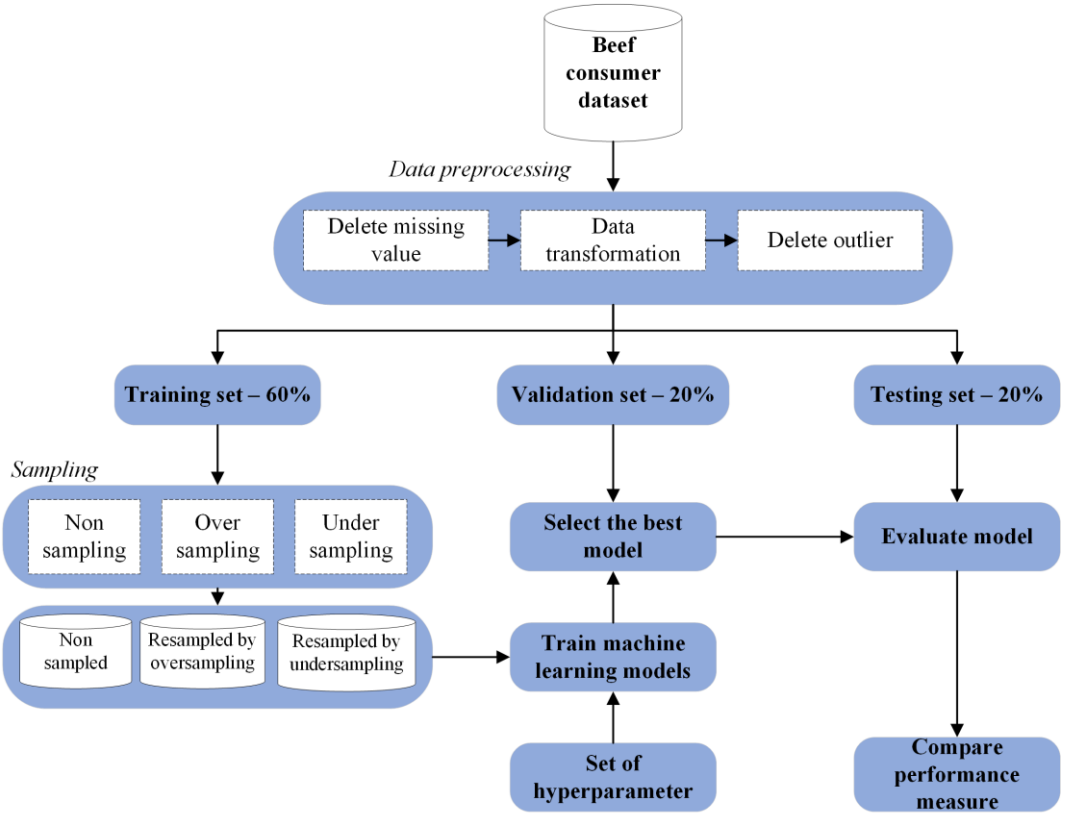


Figure 1 – Experimental design for predicting beef consumer characteristics

A resampling technique was used to address the issue of imbalance datasets. In an imbalanced dataset, the number of instances in one class (the minority class) is considerably fewer than that of another class (the majority class). A dataset is typically considered imbalanced when the proportion of the minority class is equal to or less than 10% (Wendler and Gröttrup, 2021b). This imbalance can lead to biased machine learning models when predicting membership of the minority class. Oversampling and undersampling techniques can be utilized to address the issue of imbalanced data (Kaur et al., 2019). Oversampling involves increasing the number of instances in the minority class to achieve a balanced class distribution. The objective is to generate a more representative dataset, enabling the machine learning model to learn effectively from the minority class and generate more accurate predictions. Conversely, undersampling is a data reduction method used to resolve the problem of imbalanced datasets. Undersampling entails randomly or deliberately removing instances from the majority class to equalize class distribution and create a more representative dataset. An overview of undersampling and oversampling techniques is illustrated in Figure 2.

This study conducted a comparative analysis of multiple classifiers for predicting beef consumer characteristics in Indonesia. The classifiers used in this study include Logistic

Regression (LR), Chi-squared Automatic Interaction Detector (CHAID), Random Forest (RF), artificial neural network (ANN), support vector machine (SVM) and Bayesian network (BN). A brief description of those classifiers used is given below.

LR is a widely utilized statistical technique for addressing classification and regression tasks. LR is an extension of linear regression which is utilized to assess the relationship between one or more independent variables and a binary dependent variable (Schober and Vetter, 2021). A binary variable is a category variable that can only have two distinct values or levels, such as being a beef consumer or not in the context of this study. LR models are constructed through maximum likelihood estimation, aiming to determine estimated values for the model parameters that optimize the likelihood of getting the observed dataset (Jurafsky and Martin, 2021).

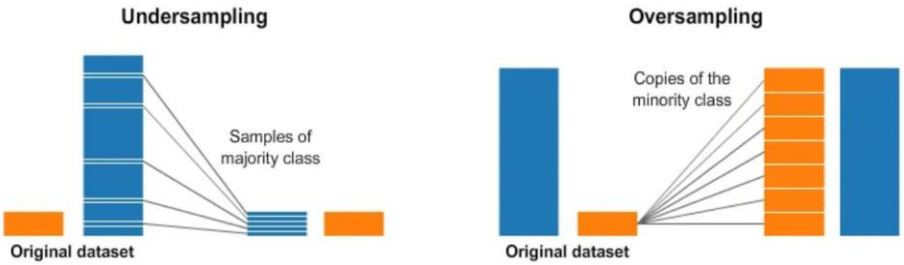


Figure 2 – Illustration for sampling technique (Raj, 2019)

CHAID statistical approach, pioneered by Kass in 1980, is widely recognized for its effective segmentation or tree growth capabilities (Wendler and Gröttrup, 2021a). This algorithm employs multi-level splits, enabling the generation of non-binary trees, wherein certain trees may possess more than two branches. Notably, CHAID accommodates all variable types and supports case weights and frequency variables. In addition, CHAID addresses missing values by considering them collectively as a single valid category (IBM, 2016).

RF algorithm is an ensemble technique that employs a parallel structured approach and leverages bagging to combine many decision tree classifiers (Liaw and Wiener, 2002). Therefore, RF is classified as an ensemble learning technique, imparting high stability and resilience against outliers. The technique is widely recognized for effectively representing intricate data structures with high precision, making it a favored approach among data miners for tackling classification challenges (Wendler and Gröttrup, 2021a).

ANN is inspired by the human brain and its functionality, which consist of multiple neurons that process and pass information between each other (Wendler and Gröttrup, 2021a). ANN is a modern computer method widely used to solve complicated real-world problems. It is particularly appealing due to their exceptional information processing capabilities, including nonlinearity, high parallelism, fault and noise tolerance, and learning and generalization capacities (Basheer and Hajmeer, 2000).

SVM is developed in the early 1990s to provide a non-linear approach for classification and regression applications (Gholami and Fakhari, 2017). SVM can offer balanced predictive performance because of their simplicity and versatility in handling various categorization issues, even with small sample sizes (Pisner and Schnyer, 2020). A SVM decision function defines an optimum hyperplane used to categorize observations into different classes based on patterns of information known as features, which can be applied to predict the most likely label for new data.

BN are structured representations of probability distributions that extend the naïve Bayesian classifier and explicitly convey information about independence. The Tree Augmented Naive Bayes (TAN) method is highlighted among these approaches for its superior performance compared to naive Bayes, while maintaining computational simplicity and resilience, typical of naive Bayes, without requiring a further search process (Friedman et al., 1997).

According to Wendler and Gröttrup (2021a), the dataset was randomly divided into 60%, 20%, and 20% for the training set, validation set and testing set, respectively, for model evaluation. Likewise, machine learning classifiers must also set up some hyperparameters to obtain the best model, as shown in Table 1.

Table 1 – The parameter set for classification models

Classifier	Parameter	Range
CHAID	Max tree depth	3,5 (default),7
LR	Classification cutoff	0.5 (default), 0.7, 0.9
RF	Number of models to build	100 (default), 200, 300
ANN	Neural network model	MLP, RBF
SVM	Kernel type	RBF, polynomial, sigmoid, linear
BN	Structure type	TAN, Markov Blanket

Babysitting, known as "'Trial and Error" or grad student descent (GSD), was used for hyper-parameter tuning to discover the best model (Yang and Shami, 2020). This technique is implemented manually and extensively utilized by students and scientists. After building a machine learning (ML) model, the study further tests numerous hyper-parameter values based on experience, guesswork, or the analysis of previously evaluated results; this process is repeated until it obtains satisfactory results. For CHAID models, the study used hyperparameters of maximum tree depth ranging from 3, 5 and 7. LR further uses classification cutoff values of 0.5, 0.7 and 0.9 to obtain optimum results. As for RF, the hyperparameter used is the number of models to build, ranging from 100, 200 and 300. Meanwhile, ANN uses MLP and RBF neural network types. In addition, SVM uses kernel types such as RBF, polynomial, sigmoid and linear for hyperparameter used. Furthermore, BN employs structure type in the form of TAN and Markov Blanket.

The confusion matrix, known as the error matrix, is one parameter to assess a classification model's performance. It illustrates the correlation between the instances of real observation and projection. Table 2 shows confusion matrix employed for the quantitative measurement of classification performance.

Table 2 – Confusion matrix

	Predictive negative	Predictive positive
Actual negative	True negative (TN)	False positive (FP)
Actual positive	False negative (FN)	True positive (TP)

Because this study focuses on the minority class (true value), the evaluation criteria used in the study include Precision and Sensitivity and F-score. Precision is defined as the ratio of true positive predictions to the total positive predictions generated by a model, while sensitivity measures the proportion of correctly predicted positive instances out of all real positive instances. F-score is a metric that takes precision and sensitivity into account. All assessments were conducted on a laptop computer with Processor Intel(R) Core (TM) i7-6500U CPU @ 2.50GHz 2.59 GHz and 16 GB RAM using Microsoft Windows 10 operating system. The data analysis was performed using SPSS Modeler 18.0.

Table 3 – The metric for data classification

Metric	Formula
Sensitivity	$\frac{TP}{TP + FN}$
Precision	$\frac{TP}{TP + FP}$
F-score	$\frac{2\ x\ sensitivity\ x\ precision}{sensitivity + precision}$

Experiment datasets were retrieved from nationwide Indonesian household socioeconomic and expenditure surveys (SUSENAS, 2021) conducted by the Central Bureau

of Statistics of the Republic of Indonesia. Based upon literature review, 14 variables are adapted in this study, namely one target variable and 13 predictors. Beef consumption acts as a target variable, stating whether respondent households have consumed beef in the past week. As for predictor variables, this study uses other food sources of protein (chicken, egg, tofu, and tempeh/ fermented soybean), expenditure as a proxy for income (Gibson and Kim, 2013), province as proxy for culture (Peri, 2004), resident location, age, education, literacy, family member, house size, and car ownership. Table 4 summarizes the variables used in the study.

Table 4 – Variables used in the study

Variables	Data type	Coding description
Target variable		
Beef consumption	Categorical	1=Consume for last week, 0=otherwise
Predictor		
Protein source food		
Chicken meat	Categorical	1=Consume for last week, 0=otherwise
Egg	Categorical	1=Consume for last week, 0=otherwise
Tofu	Categorical	1=Consume for last week, 0=otherwise
Tempeh (fermented soybean)	Categorical	1=Consume for last week, 0=otherwise
Household's socio-economic characteristics		
Total expenditure	Numerical	Average Household Expenses a Month (USD)
Province	Categorical	Indonesia has 34 provinces
Resident location	Categorical	Urban =1, rural = 0
Age	Numerical	Age of household head
Education	Categorical	No = 0, elementary school = 1, junior high school = 2, senior high school = 3, college = 4
Literacy	Categorical	Household head can read = 1, otherwise = 0
Family member	Numeric	Number of family member
House size	Numeric	Size of house (m^2)
Car ownership	Categorical	Yes = 1, otherwise = 0

A total of 33,361 respondents spread across 34 provinces in Indonesia were included in the samples. First, preprocessing the dataset was carried out, including missing value checks, data transformation, and outlier identification. Initially, 333 extreme values were found from the original dataset. Thereafter, the data was transformed using a natural logarithm (ln) to normalize the data distribution. Subsequently, there were no extreme values found in the dataset.

Table 5 – Characteristics of the respondents

Categorical variable	N	%
Education		
No school	1,403	4.21
Elementary school	14,049	42.11
Junior high school	5,645	16.92
Senior high school	8,822	26.44
College	3,442	10.32
Literacy		
Can read	31,540	94.54
Otherwise	1,821	5.46
Resident area		
Urban	13,779	41.30
Rural	19,582	58.70
Car ownership		
Have	3,532	10.59
Otherwise	29,829	89.41
Continues variable	Mean	SD
Age (years)	48.020	13.316
Ln family member	1.216	0.504
Ln house size (m^2)	4.117	0.577
Ln expenditure per month (USD)	5.485	0.627

Source: BPS [38]), processed.

The characteristics of respondents were shown in Table 5. On average, respondents are 48.02 years old, with 42.11% having attained elementary school education. The majority exhibit proficient literacy skills (94.54%) and reside in rural areas (58.70%), with an average family size of 3.78 people. In addition, respondents allocate an average of 294.011 USD per month on expenditures. Regarding assets, respondents own a house averaging 72.39 m^2 in size, and the majority (89.41%) does not own a car.

RESULTS OF STUDY

The average beef consumption per household in Indonesian is 0.021 kg per week, with a standard deviation of 0.123, ranging from 0 to 4 kg. This data reveals a significant variation in beef consumption across Indonesia. Regarding the distribution pattern, the data on beef consumption is not normally distributed (Figure 3). Instead, the data distribution exhibits asymmetry, with a positive skewness of 8.753 (positive skewness or right skewness) and kurtosis of 112.369 (characterizing a leptokurtic or heavy-tailed distribution). Positive skewness indicates that most cases fall on the left of the mean, where the mean (0.021) exceeds both the mode (0) and median (0). Leptokurtic distribution suggests a higher likelihood of the presence of outliers. Generally, skewness and kurtosis values are expected to fall within the range of +2 to -2 for a normally distributed dataset (Garson, 2012).

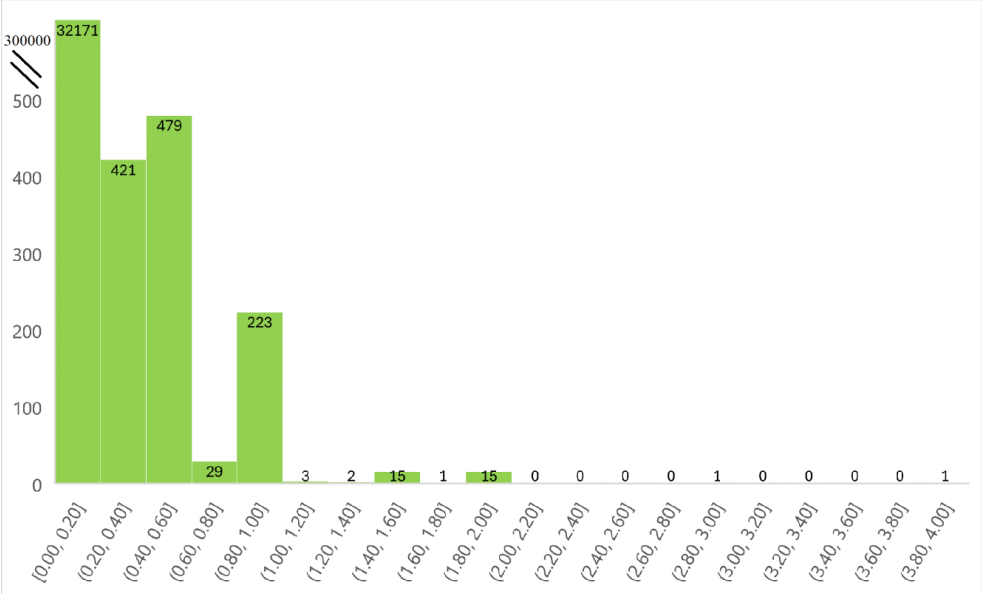


Figure 3 – The distribution of beef consumers is based on the quantity of beef consumed.

Furthermore, considering the distribution of asymmetrical data, beef consumption data needs to be transformed into binary categories, namely, consumers or non-consumers. Consequently, it reveals that the number of beef consumers in the sample was 1,351 (4.05%), whereas 32,010 (95.95%) were classified as non-beef consumers (Figure 4). The frequency data indicates a substantial disparity between the number of consumers (4.05%) and non-consumers (95.95%), indicating an imbalance dataset.

In this section, the study presents a comprehensive overview of the comparison results among machine learning techniques across three scenarios: original data (no sampling), oversampling, and undersampling. This paper further delineates the accuracy measurement into three distinct partitions: training, validation, and testing. The training dataset is utilized to train the model. The validation set is employed to optimize the model's parameters and hyperparameters. This process aids in the identification of the most optimal iteration of the model before subjecting it to evaluation using data that has not been previously encountered. The testing set, entirely separate from the training set and devoid of prior exposure to the model, serves to evaluate the model's ability to generalize effectively to new and unobserved

data instances. Consistency in the model's performance between the testing and validation sets indicates its potential for good generalization to new data. The performance of evaluation metrics is provided in Table 6, showcasing the outcomes of classifiers on both imbalanced and balanced data (oversampling and under-sampling) across all subjects.

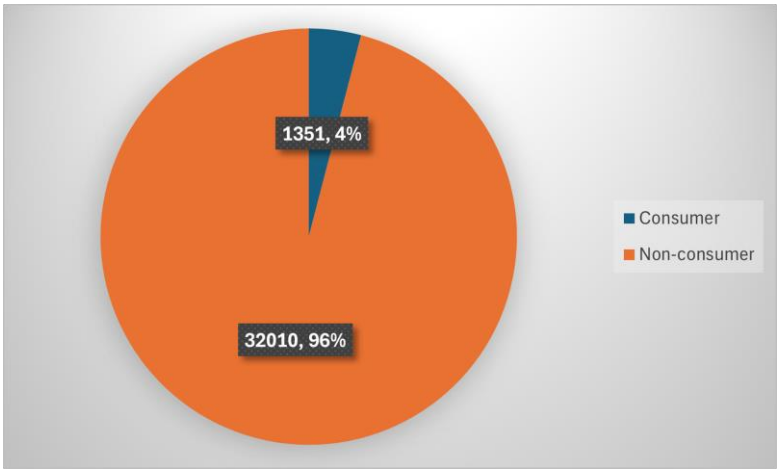


Figure 4 – Beef consumer distribution in Indonesia, 2021

Table 6 – Comparison outcomes of classifier on no sampling and resampling

Resampling method	Algorithm model	Training			Validation			Testing		
		P	S	F	P	S	F	P	S	F
No sampling	LR	0.522	0.075	0.131	0.500	0.083	0.142	0.545	0.081	0.142
	CHAID	NA	NA	NA	NA	NA	NA	NA	NA	NA
	RF	0.158	0.927	0.270	0.122	0.711	0.209	0.144	0.763	0.242
	SVM	0.753	0.270	0.397	0.271	0.135	0.180	0.316	0.146	0.200
	ANN	0.429	0.030	0.057	0.263	0.019	0.035	0.423	0.037	0.069
	BN	0.416	0.204	0.274	0.379	0.207	0.268	0.407	0.207	0.274
	Average	0.455	0.301	0.226	0.307	0.231	0.167	0.367	0.247	0.185
Oversampling	LR	0.793	0.796	0.795	0.132	0.752	0.225	0.157	0.793	0.262
	CHAID	0.790	0.813	0.801	0.120	0.711	0.205	0.145	0.756	0.243
	RF	0.876	0.970	0.920	0.150	0.613	0.241	0.183	0.661	0.287
	SVM	0.883	0.956	0.918	0.141	0.523	0.222	0.171	0.586	0.265
	ANN	0.792	0.808	0.800	0.127	0.748	0.217	0.152	0.783	0.255
	BN	0.775	0.820	0.797	0.112	0.737	0.195	0.136	0.773	0.232
	Average	0.766	0.781	0.751	0.156	0.616	0.210	0.187	0.657	0.247
Undersampling	LR	0.802	0.792	0.797	0.123	0.726	0.211	0.151	0.776	0.253
	CHAID	0.727	0.723	0.725	0.106	0.707	0.184	0.113	0.678	0.193
	RF	0.788	0.928	0.852	0.106	0.733	0.186	0.130	0.814	0.225
	SVM	0.804	0.881	0.841	0.121	0.714	0.207	0.143	0.763	0.241
	ANN	0.784	0.804	0.794	0.119	0.756	0.205	0.135	0.793	0.231
	BN	0.751	0.835	0.791	0.104	0.752	0.183	0.129	0.814	0.223
	Average	0.775	0.821	0.793	0.119	0.715	0.198	0.141	0.756	0.230

Note: P=Precision, S=Sensitivity, F=F-Score, N/A= Not applicable.

In general, resampling techniques can improve the predictive ability of classifiers for identifying beef consumers in Indonesia. In particular, oversampling yields slightly better results compared to undersampling. This is evidenced by the average F-score of the six classifiers in oversampling (testing dataset), 0.247, higher than the F-scores of undersampling (0.230) and no sampling (0.185). Moreover, there were no significant differences in precision, sensitivity, or F-score between the testing and validation datasets. This consistency underscores the validity of the classification results.

Concerning the testing set, under the imbalanced condition (no sampling), LR, RF, SVM, ANN and BN are applicable classifiers for predicting beef consumers, each exhibiting varying performances. Among them, BN demonstrates the highest F-score (0.274), followed by RF (0.242). BN displays higher precision (0.407) compared to RF (0.144). On the other hand, RF exhibits greater sensitivity (0.763) than BN (0.207). Meanwhile, CHAID is not

applicable in imbalanced data conditions, failing to predict true positives and false positives, thereby lacking precision and sensitivity.

Other than BN, all classifiers show the most optimal level of accuracy under oversampling conditions. In the oversampling condition, RF showcases superior performance with an F-score of 0.287, outperforming SVM, which achieved an F-score of 0.265. RF is more precise and sensitive than SVM.

Regarding the undersampling conditions, LR outperforms other classifiers with an F-score of 0.253, while SVM ranks second with an F-score of 0.241. LR is relatively more precise and sensitive than SVM.

In the next stage, the study compared the performance of six classifiers to identify their optimal ability. BN exhibited the best performance under the condition of no sampling, whereas the other five classifiers, LR, CHAID, RF, ANN, SVM, achieved superior results with oversampling. In terms of the predictive capability for identifying beef consumers in Indonesia, these six classifiers can be grouped into three categories based on the receiver operating characteristic curve (ROC) (Figure 5), namely classifiers with high, medium, and low accuracy. RF and SVM are categorized as classifiers with high predictive capabilities. Meanwhile, LR, CHAID and ANN fell into the medium-capability category. BN exhibited the lowest predictive capability among all six classifiers.

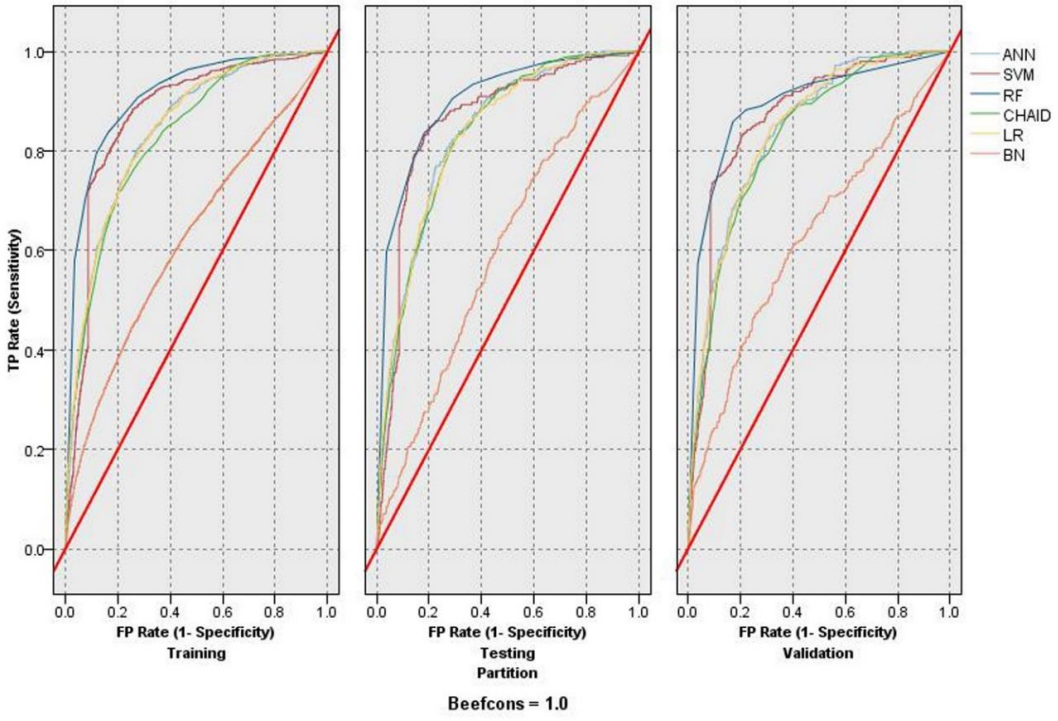


Figure 5 – Comparison of accuracy among six classifier using ROC

Table 7 shows that results of predictors' importance among six algorithms. It is evident that income level, as indicated by expenditure, exerts the most influence on predicting beef consumers, accounting for an important score of 1.5. The second most influential factor is culture, represented by province, contributing with a weight of 1.25. In addition, house size, age, education level, poultry consumption and family members play crucial roles, representing prediction powers of 0.38, 0.36, 0.36, 0.36, 0.34 and 0.30, respectively. Among the remaining predictors considered in the model, their impacts on predicting beef consumers are relatively limited. Literacy demonstrates a modest contribution, explaining 0.22 of importance. Moreover, the location of residence also affects beef consumption prediction (0.21). Furthermore, tempeh consumption (0.20), car ownership (0.19), egg consumption (0.16) and tofu consumption (0.13), also contribute to beef consumption prediction.

Table 7 – Aggregated predictor importance value

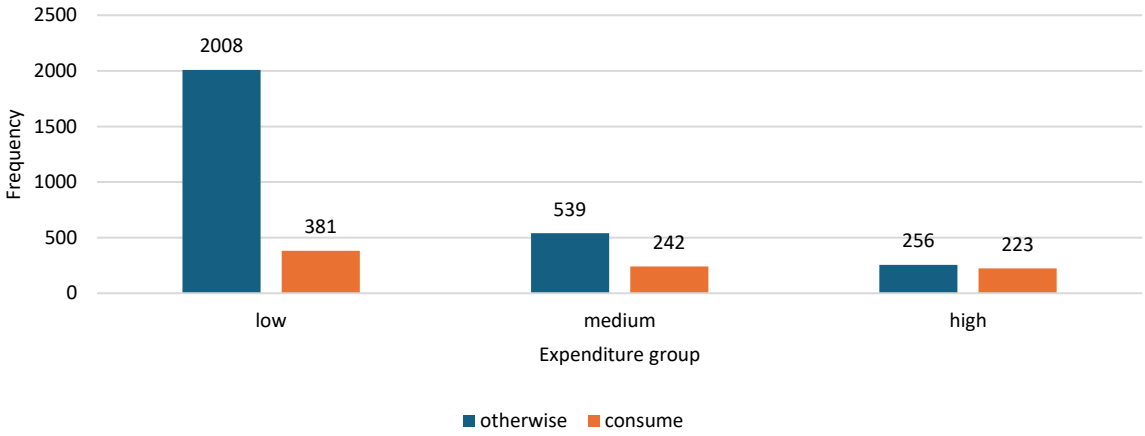
Beef consumer's predictor	Algorithm						V(fused)*
	LR	CHAID	RF	SVM	ANN	BN	
Expenditure	0.30	0.50	0.11	0.13	0.37	0.09	1.50
Province	0.29	0.17	0.24	0.26	0.20	0.09	1.25
House Size	0.03	0.06	0.12	0.03	0.06	0.08	0.38
Age	0.00	0.04	0.15	0.05	0.05	0.07	0.36
Education Level	0.04	0.05	0.12	0.00	0.07	0.08	0.36
Poultry Consumption	0.04	0.06	0.04	0.10	0.02	0.08	0.34
Family Member	0.00	0.03	0.12	0.06	0.09	0.00	0.30
Literacy	0.04	0.00	0.03	0.04	0.04	0.07	0.22
Resident Location	0.06	0.04	0.03	0.00	0.00	0.08	0.21
Tempeh Consumption	0.06	0.00	0.00	0.12	0.02	0.00	0.20
Car Ownership	0.05	0.04	0.00	0.00	0.03	0.07	0.19
Egg	0.00	0.00	0.00	0.09	0.00	0.07	0.16
Tofu Consumption	0.04	0.00	0.03	0.06	0.00	0.00	0.13

*V (Fused) *: total relative importance value for each attribute. Source: SPPS Modeler output.*

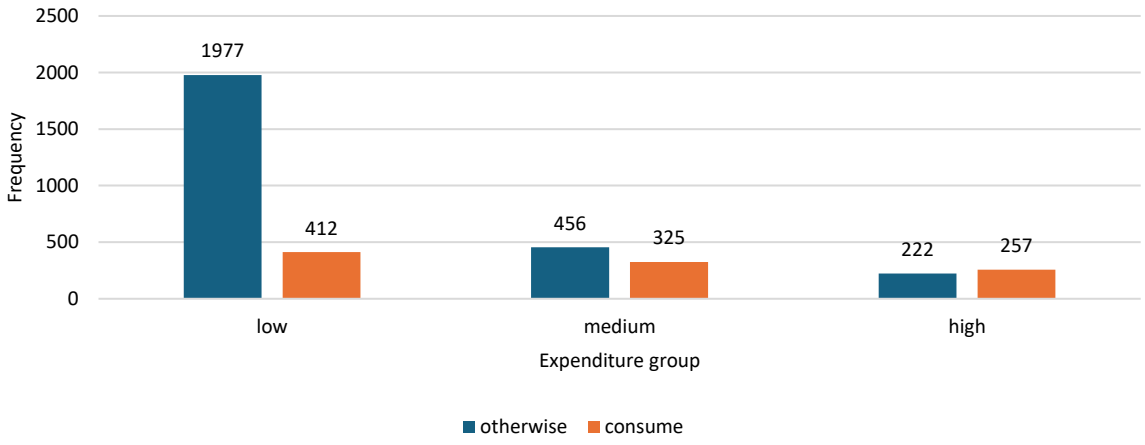
Table 8 – Result of chi square test

Classifier	Predictor	Chi-square value	df	p
SVM	Expenditure	243.841	2	0.000
	Province	1 155.926	33	0.000
RF	Expenditure	370.414	2	0.000
	Province	1 223.265	33	0.000

Source: SPSS output.



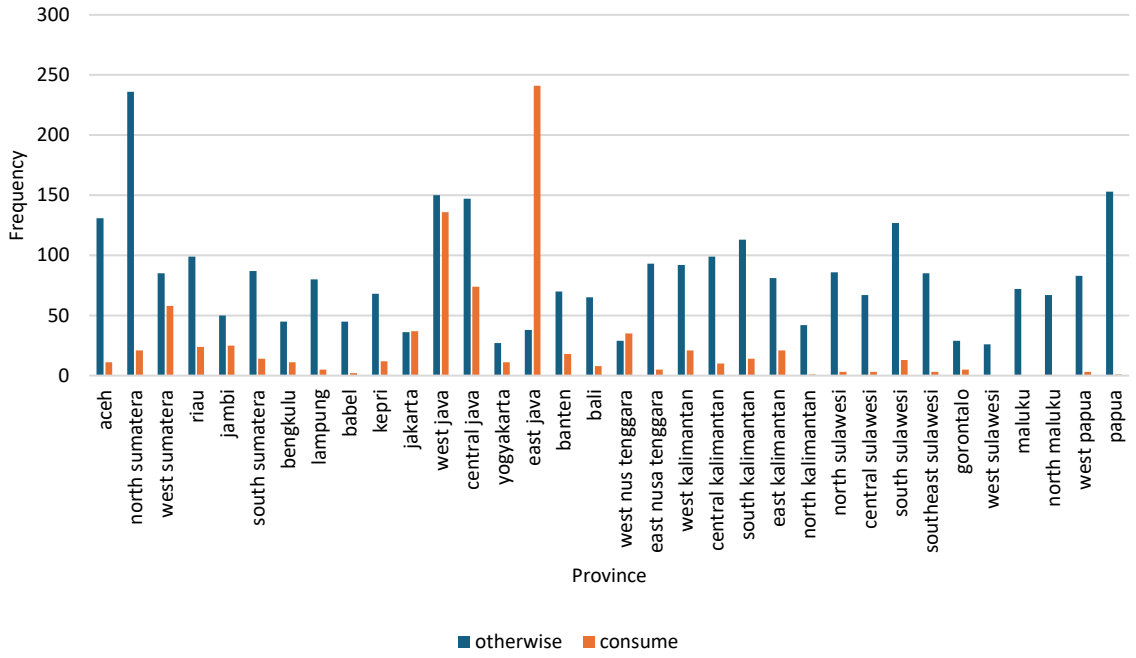
(a)



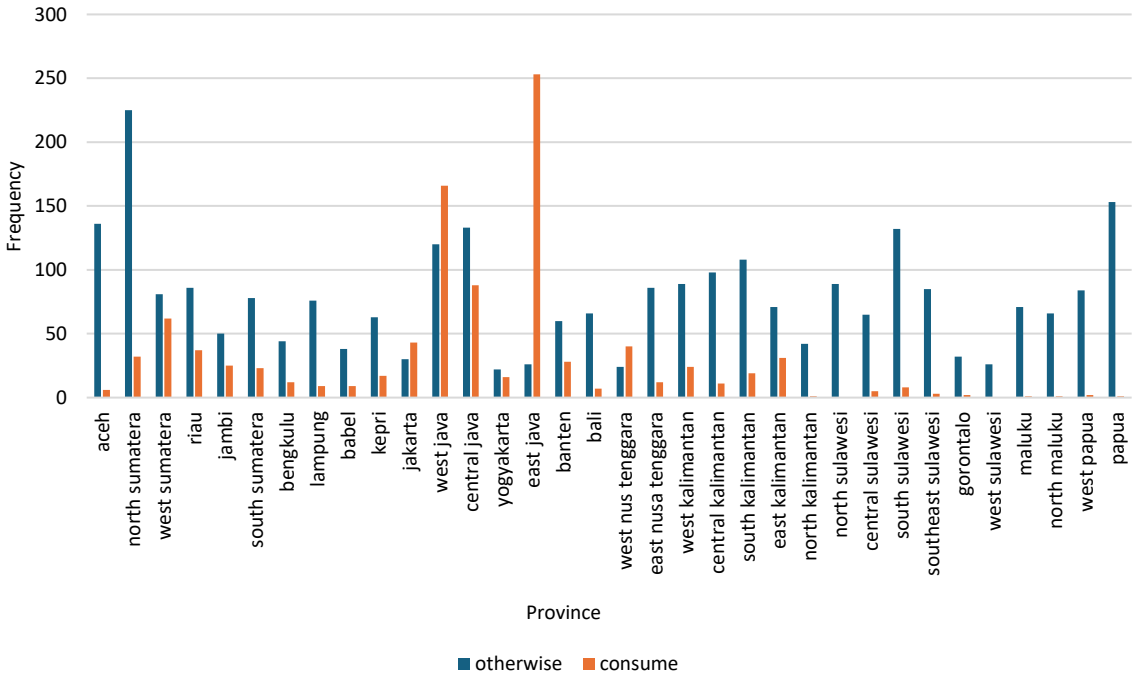
(b)

Figure 5 – Distribution of beef consumers based on expenditure, (a) SVM, (b) RF

Although SVM and RF demonstrate high predictive capabilities, interpreting their prediction results can be challenging. Therefore, this study incorporated a chi-square test to examine the relationship between predictors and prediction results in the testing dataset, especially focusing on the most important predictors (V(fused)>1). Before conducting the chi-square test, the continued data needed to be transformed into categorical data. In this case, the expenditure level was categorised into three groups based on standard values (Z scores), i.e., low (Z<0), medium (0<Z<1) and high (Z>1). The Chi-square test proves that expenditure and province have a strong influence on beef consumption, both in SVM and RF (Table 8).



(a)



(b)

Figure 6 – The distribution of beef consumer based on province, (a) SVM, (b) RF

In addition, the study also used a histogram to ease interpreting the predicted results. Based on Figure 5, it can be interpreted that higher expenditure has an impact on increasing the number of beef consumers in Indonesia. This can be seen from the ratio of consumers to non-consumers in the high group (SVM=0.871, RF=1.158) higher than the medium group (SVM=0.449, RF=0.713). Likewise, the ratio of beef consumers to non-consumers in the medium group (SVM=0.449, RF=0.713) is higher than the low group (SVM=0.190, RF=0.208).

The distribution of beef consumers by province can be seen in Figure 6. The average ratio between consumers and non-consumers is 0.402 in SVM and 0.609 in RF. Based on the ratio of the number of beef consumers to non-consumers, it can be interpreted into two groups of provinces, namely provinces with high consumption (SVM>0.402, RF>0.609) and low consumption (SVM<0.402, RF<0.609). Both SVM and RF show similar patterns about the beef consumer across province in Indonesia. Provinces that fall into the high consumption group include East Java Province, followed by West Nusa Tenggara, Jakarta, West Java, West Sumatra, Central Java, Yogyakarta, and Jambi. While the remaining provinces are included in the low consumption group, among the lowest consumption include Papua, Maluku, North Maluku, West Papua, West Sulawesi, and North Maluku.

DISCUSSION OF RESULTS

This study uses household socioeconomic and expenditure data to predict beef consumers by using machine learning algorithms combined with resampling techniques to handle an imbalanced dataset. The oversampling technique is superior to undersampling, possibly due to its ability to retain information intact, whereas undersampling may lead to a loss of important data. However, oversampling has two major drawbacks, including increased potential for overfitting and longer computation time due to the increased sample size (Mohammed et al., 2020).

The performance of six algorithms is subsequently evaluated using established performance metrics, with RF and SVM exhibiting superior predictive performance when combined with oversampling techniques for beef consumption prediction. RF's effectiveness in classification with high accuracy stems from ensemble learning (Dong et al., 2020), resistance to noise (Dong et al., 2020), and capacity to recognise nonlinear relationships. Similarly, SVM demonstrates robust classification capabilities attributed to factors such as effectiveness in high dimensional spaces, robustness to overfitting, versatility in kernel selection, and flexibility in handling linear and non-linear data with controlled complexity (Pisner and Schnyer, 2020).

The key finding underscores the linear and positive influence of income level on beef consumption. This could be attributed to the fact that beef is perceived as a luxury food item in Indonesia [6], thereby limiting its accessibility to high-income households. Furthermore, this study emphasises income level as the predominant determinant predicting beef consumption, which is consistent with previous studies conducted by Uzunöz and Karakaş (2014), Jabo and Zaharadden (2018), Popoola et al. (2022). Given beef is relatively more expensive in Indonesia, this finding aligns with economic theory, which postulates that higher prices lead to lower the demand for a good. This conclusion is reinforced by empirical evidence reflecting higher demand for chicken and eggs compared to beef (Ministry of Agriculture of Indonesia, 2022).

The province serves as a representation of culture, constituting the second determining factor affecting beef consumption among individuals. The provinces of East Java, West Nusa Tenggara, Jakarta, West Sumatra, and West Java display the highest levels of beef consumption in Indonesia. In contrast, provinces in eastern Indonesia demonstrate lower levels of consumption. These consumption patterns of beef may be influenced by various factors, including religious beliefs, traditional customs, or culinary inclinations, leading to variations in dietary habits. Understanding the cultural intricacies of ethnic groups is essential, often requiring a comprehensive awareness of their rituals and practices concerning food (Nam et al., 2010). For instance, West Sumatra boasts a culinary tradition

centred around beef, with rendang-a renowned dish in Indonesia-serving as a prominent specialty (Fatimah et al., 2021). On the other hand, in Sulawesi, particularly among the Toraja ethnic group, cattle hold significant symbolic value as sacred animals often served in traditional ceremonies, such as funeral ceremonies (Baan et al., 2022). In addition, dietary patterns in eastern Indonesia often prioritise alternative protein sources such as fish and seafood (Soselisa et al., 2021). Conversely, the Minahasa ethnic group in North Sulawesi exhibits distinctive culinary customs, including the consumption of dog and rat meat, with beef being less popular (Laatung et al., 2019). Such unique dietary practices shape individuals' preferences regarding beef consumption.

This study is confined to certain limitations. The sampling technique used in this study still relies on basic methods and does not explore more sophisticated methods. In addition, the number of classifiers used is limited to six. Future research is recommended to explore more advanced sampling techniques and incorporate additional classifiers to enhance predictive accuracy.

CONCLUSION

To achieve the objectives, a classification model is used to analyse datasets extracted from expenditure and socioeconomic data of Indonesia. Building on previous research, this study expands the methodological framework by combining several machine learning algorithms with resampling techniques, providing methodological advancements. The findings confirm and extend prior research results, highlighting significant disparities in beef consumption across diverse socioeconomic contexts. Specifically, this study confirms that income level, cultural practices, house size, education level, age, and consumption of poultry meat significantly influence beef consumption patterns.

The results of this study hold practical implications and can be effectively utilised by marketing professionals and policymakers. By considering the income level and cultural practices prevalent in different regions, stakeholders in beef industry can strategize and adapt their approaches to increase demand and marketing efficiency. For example, the allocation of supply and pricing strategies may vary according to the income level of different areas. Similarly, decisions regarding the number and location of sales outlets should take into consideration local economic and cultural factors. For policymakers, the results of this research can serve as a valuable reference for determining areas for cattle farming development initiatives.

An additional implication of this study lies in the development of a methodology tailored to address imbalanced datasets, complemented with the application of machine learning algorithms to predict beef consumers. In contrast to previous studies, which primarily relied on traditional methods such as linear regressions and analysis of variance (ANOVA), this study introduces a novel approach that expands the methods for predicting consumer consumption patterns. Therefore, it is plausible that this novel methodology could extend beyond beef consumption prediction in this context, offering applicability to other product/service categories and facilitating consumer segmentation efforts.

DATA AVAILABILITY STATEMENT AND ACKNOWLEDGEMENTS

Data will be made available on request. The first author would like thanks to Taiwan ICDF for providing scholarship in study. First author also is thankful to Ministry of Agriculture of Indonesia for supporting the research and study.

REFERENCES

1. Abbasi, B., Goldenholz, D.M., 2019. Machine learning applications in epilepsy. Epilepsia 60, 2037–2047. https://doi.org/10.1111/epi.16333.
2. Agus, A., Widi, T.S.M., 2018. Current situation and future prospects for beef cattle production in Indonesia - A review. Asian-Australasian J. Anim. Sci. 31, 976–983.

https://doi.org/10.5713/ajas.18.0233.
3. Anindita, R., Sadiyah, A.A., Khoiriyah, N., Nendyssa, D.R., 2020. The demand for beef in Indonesian urban. IOP Conf. Ser. Earth Environ. Sci. 411, 1–9. https://doi.org/10.1088/1755-1315/411/1/012057.
4. Baan, A., Girik Allo, M.D., Patak, A.A., 2022. The cultural attitudes of a funeral ritual discourse in the indigenous Torajan, Indonesia. Heliyon 8, e08925. https://doi.org/10.1016/j.heliyon.2022.e08925.
5. Basheer, I.A., Hajmeer, M., 2000. Artificial neural networks: fundamentals, computing, design, and application. J. Microbiol. Methods 43, 3–31. https://doi.org/https://doi.org/10.1016/S0167-7012(00)00201-3.
6. BPS, 2021. Expenditure for consumption of Indonesia based on March 2021 SUSENAS. Badan Pusat Statistik Indonesia, Jakarta.
7. Cuocolo, R., Cipullo, M.B., Stanzione, A., Ugga, L., Romeo, V., Radice, L., Brunetti, A., Imbriaco, M., 2019. Machine learning applications in prostate cancer magnetic resonance imaging. Eur. Radiol. Exp. 3, 35. https://doi.org/10.1186/s41747-019-0109-2.
8. Davagdorj, K., Lee, J.S., Pham, V.H., Ryu, K.H., 2020. A comparative analysis of machine learning methods for class imbalance in a smoking cessation intervention. Appl. Sci. 10, 3307. https://doi.org/10.3390/app10093307.
9. Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q., 2020. A survey on ensemble learning. Front. Comput. Sci. 14, 241–258. https://doi.org/10.1007/s11704-019-8208-z.
10. Fatimah, S., Syafrini, D., Wasino, Zainul, R., 2021. Rendang lokan: history, symbol of cultural identity, and food adaptation of Minangkabau tribe in West Sumatra, Indonesia. J. Ethn. Foods 8, 1–10. https://doi.org/10.1186/s42779-021-00088-2.
11. Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian Network Classifier. Mach. Learn. 29, 131–163. https://doi.org/https://doi.org/10.1023/A:1007465528199.
12. Garaus, M., Garaus, C., 2023. US consumers' mental associations with meat substitute products. Front. Nutr. 10, 1–11. https://doi.org/10.3389/fnut.2023.1135476.
13. Garson, G.D., 2012. Testing statistical assumptions, Statistical Associate Publishing. Aheboro, NC 27205 USA.
14. Gholami, R., Fakhari, N., 2017. Support Vector Machine: Principles, Parameters, and Applications, in: Handbook of Neural Computation. Elsevier Inc., pp. 515–535. https://doi.org/10.1016/B978-0-12-811318-9.00027-2.
15. Gibson, J., Kim, B., 2013. How reliable are household expenditures as a proxy for permanent income? Implications for the income-nutrition relationship. Econ. Lett. 118, 23–25. https://doi.org/10.1016/j.econlet.2012.09.016.
16. Hadi, S.N., Chung, R.H., 2022. Estimation of Demand for Beef Imports in Indonesia : An Autoregressive Distributed Lag (ARDL) Approach. Agric. 12, 1212. https://doi.org/10.3390/agriculture12081212.
17. Hamdan, A., Sumantri, I., Hadi, S.N., Rohaeni, E.S., Yanti, N.D., Chang, C., 2019. A market chain analysis of interisland cattle trade into South Kalimantan, Indonesia. IOP Conf. Ser. Earth Environ. Sci. 387, 1–6. https://doi.org/10.1088/1755-1315/387/1/012038
18. Hovhannisyan, V., Devadoss, S., 2020. Effects of urbanization on food demand in China. Empir. Econ. 58, 699–721. https://doi.org/10.1007/s00181-018-1526-4.
19. IBM, 2016. IBM SPSS Modeler 18.0 Algorithms Guide. IBM, Chicago.
20. Jabo, M.S.M., Zaharadden, I.M., 2018. Estimation of beef consumption : An application of econometric model in Wamakko local government area of Sokoto State , Nigeria. Direct Res. J. Agric. Food Sci. 6, 84–88. https://doi.org/https://doi.org/10.26765/DRJAFS.2018.3107.
21. Jurafsky, D., Martin, J., 2021. Logistic regression. Encycl. Qual. Life Well-Being Res. https://doi.org/https://doi.org/10.1007/978-3-319-69909-7_1689-2.
22. Kaur, H., Pannu, H.S., Malhi, A.K., 2019. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Comput. Surv. 52, 1–36. https://doi.org/10.1145/3343440.
23. Khoiriyah, N., Anindita, R., Hanani, N., Wahib Muhaimin, A., 2020. Impacts of Rising Animal Food Prices on Demand and Poverty in Indonesia. Agric. Soc. Econ. J. 20, 67–

78. https://doi.org/10.21776/ub.agrise.2020.20.1.9.
24. Kwekha-Rashid, A.S., Abduljabbar, H.N., Alhayani, B., 2023. Coronavirus disease (COVID-19) cases analysis using machine-learning applications. Appl. Nanosci. 13, 2013–2025. https://doi.org/10.1007/s13204-021-01868-7.
25. Laatung, S., Fuah, A.M., Masy'Ud, B., Sumantri, C., Dohong, S., 2019. The Hunting of White-Tailed rat by Minahasa Tribe, North Sulawesi, Indonesia. IOP Conf. Ser. Earth Environ. Sci. 399, 012032. https://doi.org/10.1088/1755-1315/399/1/012032.
26. Liaw, A., Wiener, M., 2002. The R Journal: Classification and regression by randomForest. R J. 2, 18–22.
27. Ministry of Agriculture of Indonesia, 2022. Livestock and Animal Health Statistics 2022. Directorate General of Livestock and Animal Health Services, Jakarta.
28. Mohammed, R., Rawashdeh, J., Abdullah, M., 2020. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results, in: 2020 11th International Conference on Information and Communication Systems (ICICS). IEEE, Irbid, Jordan, pp. 243–248. https://doi.org/10.1109/ICICS49469.2020.239556.
29. Nam, K.C., Jo, C., Lee, M., 2010. Meat products and consumption culture in the East. Meat Sci. 86, 95–102. https://doi.org/10.1016/j.meatsci.2010.04.026.
30. Peri, G., 2004. Socio-Cultural Variables and Economic Success : Evidence from Italian 1951-1991. J. Contrib. Macroecon. 4, 1–34.
31. PIHPS, 2023. Informasi Harga Pangan Antar Daerah [WWW Document]. URL https://www.bi.go.id/hargapangan (accessed 8.2.23).
32. Pisner, D.A., Schnyer, D.M., 2020. Support vector machine, in: Machine Learning: Methods and Applications to Brain Disorders. Elsevier Inc., pp. 101–121. https://doi.org/10.1016/B978-0-12-815739-8.00006-7.
33. Popoola, I.O., Soladoye, P.O., Gaudette, N.J., Wismer, W. V., 2022. A Review of Sensory and Consumer-related Factors Influencing the Acceptance of Red Meats from Alternative Animal Species. Food Rev. Int. 38, 266–285. https://doi.org/10.1080/87559129.2020.1860084.
34. Raj, J., 2019. What To Do When Your Classification Data is Imbalanced [WWW Document]. URL https://towardsdatascience.com/what-to-do-when-your-classification-dataset-is-imbalanced-6af031b12a36 (accessed 7.27.23).
35. Sann, R., Lai, P.C., Liaw, S.Y., Chen, C.T., 2022. Predicting Online Complaining Behavior in the Hospitality Industry: Application of Big Data Analytics to Online Reviews. Sustain. 14, 1800. https://doi.org/10.3390/su14031800.
36. Schober, P., Vetter, T.R., 2021. Logistic Regression in Medical Research. Int. Anesth. Res. Soc. 132, 365–366.
37. Sharma, A., Jain, A., Gupta, P., Chowdary, V., 2021. Machine Learning Applications for Precision Agriculture: A Comprehensive Review. IEEE Access 9, 4843–4873. https://doi.org/10.1109/ACCESS.2020.3048415.
38. Sharma, R., Kamble, S.S., Gunasekaran, A., Kumar, V., Kumar, A., 2020. A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. Comput. Oper. Res. 119, 104926. https://doi.org/10.1016/j.cor.2020.104926.
39. Soselisa, P.S., Soselisa, H.L., Alfons, C.R., 2021. Seafood consumption patterns among coastal people in Ambon Island, Eastern Indonesia. Food Res. 5, 1–6. https://doi.org/10.26656/fr.2017.5(S3).005.
40. Townsend, J., Micucci, C.P., Hymel, J.H., Maroulas, V., Vogiatzis, K.D., 2020. Representation of molecular structures with persistent homology for machine learning applications in chemistry. Nat. Commun. 11, 1–9. https://doi.org/10.1038/s41467-020-17035-5.
41. Uzunöz, M., Karakaş, G., 2014. Socio-economic Determinants of Red Meat Consumption in Turkey: A Case Study. Çankırı Karatekin Üniversitesi Sos. Bilim. Enstitüsü Derg. 5, 37–52.
42. Wendler, T., Gröttrup, S., 2021a. Classification models, in: Data Mining with SPSS

Modeler. Springer Nature, Switzerland, pp. 753–1088. https://doi.org/https://doi.org/10.1007/978-3-030-54338-9.
43. Wendler, T., Gröttrup, S., 2021b. Imbalanced Data and Resampling Techniques, in: Data Mining with SPSS Modeler. Springer Nature, pp. 1147–1191. https://doi.org/10.1007/978-3-030-54338-9_10.
44. Yang, L., Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing 415, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061.
45. Yildirim, I., Ceylan, M., 2008. Urban and rural households' fresh chicken meat consumption behaviors in Turkey. Nutr. Food Sci. 38, 154–163. https://doi.org/doi.org/10.1108/00346650810863037.
46. Zhang, H., Wang, J., Martin, W., 2018. Factors affecting households' meat purchase and future meat consumption changes in China: a demand system approach. J. Ethn. Foods 5, 24–32. https://doi.org/10.1016/j.jef.2017.12.004.