International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

Predicting the Filipino Household Income Using Naive Bayes Classification Algorithm

Jhumer O. Apus[1], Kurt Donn V. Mantalaba[1], Abdul Jabbar B. Mackno[1] and Paul B. Bokingkito Jr.[1]

[1] Mindanao State University-Iligan Institute of Technology, Iligan City, Philippines

jhumer.apus@g.msuiit.edu.ph, kurtdonn.mantalaba@g.msuiit.edu.ph, abduljabbar.mackno@g.msuiit.edu.ph, pauljr.bokingkito@g.msuiit.edu.ph

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: The Family Income and Expenditure Survey (FIES) in the Philippines provides crucial data on household income and expenses. This study utilizes the Naive Bayes algorithm to predict Filipino economic class using household expenditure and income variables. The research aims to contribute to poverty reduction efforts by providing a predictive model for identifying vulnerable households and designing appropriate interventions. Data preprocessing steps, including data cleaning, transformation, and analysis, were performed before feature selection and modelling. Predictive models using Naive Bayes were evaluated and validated, with accuracy measured using a confusion matrix. Results show high accuracy rates, with bagging (93%) and boosting (89%) ensemble techniques used for model implementation. Findings can potentially support local government units in poverty reduction programs and policymaking. Future research could explore other machine learning algorithms and consider additional parameters to further improve model accuracy using the increasing data from the FIES datasets provided by the Philippine Statistics Authority.

Keywords: Naïve Bayes Algorithm, Household Income, Classification, Machine Learning

1. INTRODUCTION

In the Philippines, parenting costs have significant implications for family finances, with larger families spending a higher percentage of their income on raising children. Over 33% of households have seven or more family members, particularly in rural areas, increasing the risk of poverty [1]. Families with more children tend to allocate fewer resources per child, with food expenditure accounting for a substantial portion of household expenses [2]. The 2018 Family Income and Expenditure Survey (FIES) is a crucial data source for assessing poverty in the Philippines conducted every three years by the National Statistics Office, providing insights into household spending patterns [3]. In 2018, food expenditure alone accounted for 41.5% of total household expenses, much higher than in Western countries [4]. These statistics underscore the critical role of data sources like FIES in poverty projections and policy-making efforts. The survey collects data on various aspects of house-hold income and expenditure, including consumption levels, sources of income and spending patterns across different socioeconomic categories. It also includes information on the number of family members employed, their occupation, age, and educational attainment, as well as other housing characteristics. The data gathered through the FIES is critical to identifying areas of poverty and understanding the effects of economic policies on households of varying income levels.

Several studies have been conducted utilizing different machine learning techniques, including decision trees and neural networks, for predicting Filipino economic class [5,6,7]. However, most of these studies have used socio-demographic variables such as age, education level, and occupation as input features for their models. The lack of exploration of the potential of using other variables, such as household expenditure and income, has been identified in the literature. In this study, Naive Bayes algorithm was used with a focus on household expenditure and income variables as input features. One significant advantage of the Naive Bayes algorithm is its simplicity and speed, allowing for efficient processing of large amounts of data with high accuracy while requiring minimal computational resources. In addition, Naive Bayes is robust to irrelevant features, noisy data, and missing values, making it a suitable choice for real-world data analysis. Naive Bayes is also preferred over other classification algorithms because of its ability to handle both discrete and continuous data

E-mail:author's email

types, which is a desirable feature for many classification tasks [6]. Furthermore, Naive Bayes can handle multi class problems with ease, where other classifiers may require additional modifications to handle such tasks.

This paper aims to predict Filipino economic class using Naive Bayes, a machine learning classification algorithm with FIES dataset to develop a model for predicting Filipino income classes. By leveraging machine learning algorithms, this research aims to contribute to poverty reduction efforts in the Philippines by providing a predictive model that can assist policymakers in identifying vulnerable households and designing appropriate interventions.

2. METHODOLOGY

A. Data Preparation

Data preparation is a crucial step in the machine learning pipeline that involves cleaning, transforming, and integrating data to ensure its suitability for model training. Proper data preparation is essential to enhance the efficiency and effectiveness of machine learning models, as the quality of input data greatly influences model performance. The FIES dataset was acquired from the National Statistics Office in a CSV file format. The CSV file contained raw data with various sample attributes, including total household income, region, total food expenditures, main source of income, and agricultural household. Table I presents an overview of the attributes in the raw data.

TABLE I. SAMPLE FIES RAW DATA.

Total Household Income	***Region***	***Total Food Expenditure***	***Main Source of Income***	***Agricultural Household Indicator***
XX	XX	XX	XX	XX
XX	XX	XX	XX	XX
XX	XX	XX	XX	XX

In this study, each instance of households in the dataset was labeled and classified based on their total income, utilizing the standard Income bracket of Philippine household income classes depicted in Table II. The total household income in the dataset was originally defined on an annual basis but was converted into a monthly equivalent by dividing it by 12.

TABLE II. STANDARD INCOME BRACKET OF PHILIPPINE HOUSEHOLDS.

Monthly Income Range	***Income Class***
P0 - P10,957	Poor
P10,957 - P21,914	Low-income but not poor
P21,914 - P43,828	Lower-middle
P43,828 - P76,669	Middle
P76,669 - P131,484	Upper-middle
P131,484 - P219,140	Upper-middle but not rich
P219,140-above	Rich

The raw data from the CSV file required thorough data preparation to ensure its quality and suitability for model training. In this study, the researcher used data cleaning techniques to fix or remove incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within the dataset. Columns with standard missing values such as blank cells, NaN, n/a and non-standard missing values such as "..", " ", "—" , "missing", "na", "@", "??", "***", etc. had been filled up. For handling missing and outliers, the researchers used the measure of central tendency. The mode is used to fill up the missing values for categorical attributes [9]. For columns that are numeric, the researchers used the mean and median to fill the missing values [10]. After the data preparation process, the cleaned data was imported and stored in a PostgreSQL database for further analysis and processing. PostgreSQL is a widely used open-source relational database management system known for its scalability, performance, and data integrity features, making it suitable for storing large datasets.

B. Feature Selection

Feature selection is conducted to isolate the most consistent, non-redundant, and relevant features to use in model construction [11]. As the number and variety of datasets increase, it is more important to methodically reduce them. The fundamen-tal purpose of selecting features is to improve predictive model performance while decreasing modeling costs [12] [13]. Using univariate statistical tests, the researchers identify features that have minimal or no correlation to the dependent variable.

The test compares each feature to the target variable to check if there is a statistically significant relationship between the two. Table III shows the selected features and those that yielded low scores from the univariate statistical test were eliminated. Also, features with the highest scores are selected [14]. All features have undergone univariate statistical tests. The features with higher scores have been selected, described as an independent variable. The "Income Class" features have been the dependent variable since it is the target variable for classification.

TABLE III. SELECTED FEATURES AND SCORES

Features	*Scores*
Total Income from Entrepreneurial Activities	5.46
Housing and water Expenditure	1.44
Imputed House Rental Value	1.03
Total Food Expenditure	7.23
Transportation Expenditure	5.54
Education Expenditure	5.23
Miscellaneous Goods and Services Expenditure	4.96
Restaurant and hotels Expenditure	4.71
Medical Care Expenditure	4.06
Communication Expenditure	2.92
Special Occasions Expenditure	2.27
Clothing, Footwear and Other Wear Expenditure	1.78

The selected features used for the training are "Total Income from Entrepreneurial Activities" with the highest score of 5.46, "Housing and water Expenditure", "Imputed House Rental Value", "Total Food Expenditure", "Transportation Expenditure", "Education Expenditure", "Miscellaneous Goods and Services Expenditure", "Restaurants and hotels Expenditure", "Medical Care Expenditure", "Communication Expenditure", "Special Occasions Expenditure", "Clothing, Footwear and Other Wear Expenditure", and "Crop Farming and Gardening expenses".

C. Naive Bayes Model Implementation

After data preparation and feature selection, the dataset was then ready to be fed into the Naïve Bayes Algorithm. First, the dataset is divided into two parts: the training set and the testing set. A training dataset is a large dataset used to train a machine-learning model. In machine learning, a test set is a secondary (or tertiary) data set that is used to test a machine learning algorithm after it has been trained on an initial training data set [15]. The researchers used the 80-20 rule, also known as the Pareto Principle which yields good results in several studies [16] [17] [18]. For best results when classifying the values, the researchers have split the data, 80% for the training set and 20% for the testing set.

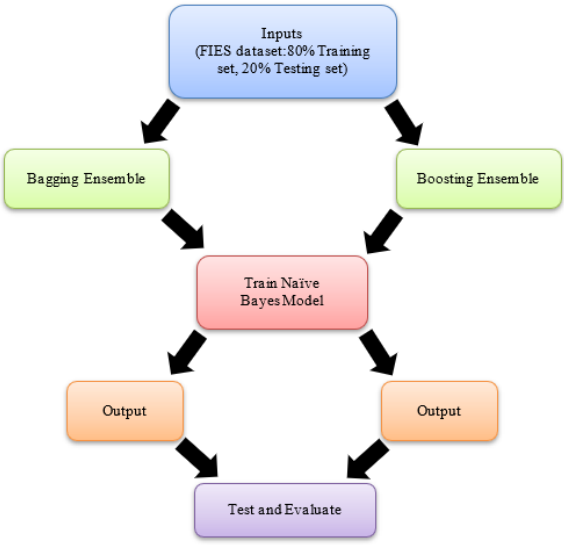

Figure 1. Architecture for model design

As shown in Figure 1, the researchers used two ensemble techniques for training Naive Bayes models, Bagging and Boosting. Bagging is a way to decrease the variance in the prediction by generating additional data for training from the dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique that adjusts the weight of an observation based on the last classification [19]. The two models are tested and evaluated according to the model's precision, recall, and accuracy.

The researchers used the python programming language to implement the Naïve Bayes algorithm techniques. The machine-learning python libraries used are Pandas, Numpy, and sklearn. Pandas provides high-performance, easy-to-use data structures, and data analysis tools for the labeled data. Sklearn provides a range of supervised and unsupervised learning algorithms such as SVMs, random forests, k-means clustering, and Naïve Bayes. NumPy adds support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays [20].

D. Model Evaluation and Validation

The predictive capability of the Naïve Bayes models for household income was then tested and validated by comparing the results of the predicted income class to the actual income class. The evaluation of model architectures was conducted to determine which model yielded higher accuracy. Furthermore, a confusion matrix was used to show the combination of the actual and predicted outputs. The instances in a predicted class are represented by the rows of the matrix, while the instances in an actual class are represented by the columns. This is a useful way to measure whether the models can account for class property overlap and identify which classes are most often confused. As shown in Figure 2, the actual data labeled on the y-axis describes the true value in the dataset and the predicted data labeled on the x-axis describes the predicted value in the dataset. The matrix shows 7 categories and the value is defined as P as Poor, LIP as Low-income but not poor, LM as Lower-middle, M as middle, UM as Upper-middle, UMR as Upper-middle but not rich, and R as rich. To determine the correct prediction the value in the actual data should be the same as in the predicted data.

Predicted Data

Actual Data	Poor	Low-income but not poor	Lower-middle	Middle	Upper-middle	Upper-middle but not rich	Rich
Poor	**P-P**	P-LIP	P-LM	P-M	P-UM	P-UMR	P-R
Low-income but not poor	LIP-P	**LIP-LIP**	LIP-LM	LIP-M	LIP-UM	LIP-UMR	LIP-R
Lower-middle	LM-P	LM-LIP	**LM-LM**	LM-M	LM-UM	LM-UMR	LM-R
Middle	M-P	M-LIP	M-LM	**M-M**	M-UM	M-UMR	M-R
Upper-middle	UM-P	UM-LIP	UM-LM	UM-M	**UM-UM**	UM-UMR	UM-R
Upper-middle but not rich	UMR- P	UMR-LIP	UMR-L M	UMR-M	UMR-U M	**UMR-UMR**	UMR-R
Rich	R-P	R-LIP	R-LM	R-M	R-UM	R-UMR	**R-R**

Figure 2. Confusion Matrix for evaluating the model.

The diagonal data cells of the confusion matrix table shows the correct prediction. The total number of correction predictions of each category starts from the top-left of the matrix diagonally to the bottom-right of the matrix. The results provide a classification report that determines the precision, recall, and f1-score of a model. Precision is one measure of a machine learning model's performance, the accuracy of a model's positive prediction. The number of true positives divided by the total number of positive predictions is known as precision [21]. Precision is defined as follows:

$$Precision = \frac{True\ Positive(TP)}{True\ Postivie(TP)+False\ Positive(FP)} \quad (1)$$

The recall is the measure of the model correctly identifying True Positives. Recall is a statistic that measures the number of positive predictions were made out of all possible positive predictions [21]. Recall is defined as follows:

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP)+False\ Negative(FN)} \quad (2)$$

F1 Score is the weighted average of Precision and Recall. F1-score is one of the most important evaluation metrics in machine learning. It elegantly sums up the predictive performance of a model by combining two otherwise competing metrics precision and recall [21]. F1 score defined as follows:

$$F1\ score = \frac{Precision*Recall}{Precision+Recall} \quad (3)$$

To determine the accuracy of the model, the total number of correct predictions were divided by the total of all predictions. The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is used along with precision and recall, which are other metrics that use various ratios of true/false positives/negatives [22]. The accuracy is calculated by the total number of correct predictions (TP + TN) divided by the total number of a dataset (P + N).

$$Accuracy = \frac{correct\ prediction\ (TP+TN)}{all\ predictions\ (P+N)} \quad (4)$$

3. Results and Discussion

A. FIES Raw Data Preparation

The acquired FIES dataset includes households in the different regions in the Philippines. It contains 41,545 instances and 60 features in a csv file format. Several anomalies are found in the dataset. Missing, empty and outliers are replaced accordingly. Table 3 shows the result of data correction, N/A values found in the dataset was replaced with “Rice Farmer” using the mode of the particular attribute. Table IV shows the result of the sample state before and after data cleaning. On this sample N/A is replaced with “Rice Farmer” using the mode formula.

TABLE IV. Sample Data of Before and After Data Cleaning

HOUSEHOLD HEAD OCCUPATION FEATURE	
BEFORE	***AFTER***
Rice Farmers	Rice Farmers

HOUSEHOLD HEAD OCCUPATION FEATURE	
BEFORE	***AFTER***
N/A	Rice Farmers
N/A	Rice Farmers
N/A	Rice Farmers
Rice Farmers	Rice Farmers
Rice Farmers	Rice Farmers
Rice Farmers	Rice Farmers

Table IV shows the result of the sample state before and after data transformation. On this sample, ”Farmhands and laborers” and “Rice farmers” replaced with numeric values 0 and 1.

TABLE V. SAMPLE DATA OF BEFORE AND AFTER DATA TRANSFORMING

HOUSEHOLD HEAD OCCUPATION FEATURE	
BEFORE	***AFTER***
Farm and laborers	0
Rice Farmers	1
Rice Farmers	1
Rice Farmers	1
Rice Farmers	1
Farm and laborers	0
Farm and laborers	0

B. Selected Features

After the data correction and data transformation, the researchers conducted the feature selection process using univariate feature selection of the 60 features of FIES dataset. The function that used for the selection process is the SelectKBest function from sklearn library. The SelectKBest function removes all the features except the top specified numbers of features. The *k* value was used to keep the top features with the high scores. The score_function of chi-squared is chosen to measure dependence between stochastic variables; this function removes the features that are the most likely to be independent of class and therefore irrelevant for classification [23]. The result of the feature selection process using univariate selection yielded 13 features with higher scores which correspond stronger correlation to the dependent variable.

C. Model Evaluation and Results

The predictive performance of Filipino Household income of Naïve Bayes were assessed using two (2) ensemble techniques, bagging and boosting ensembles. Model I is implemented using bagging ensemble and Model II using boosting ensemble. The models were implemented with 13 selected features to identify model with better accuracy. Figure 2 shows the confusion matrix report of Model I and Model II. Poor Income class yields the most number of True Positives result while Rich Income class have the least true positives for both models since it also have the least number of income class instances.

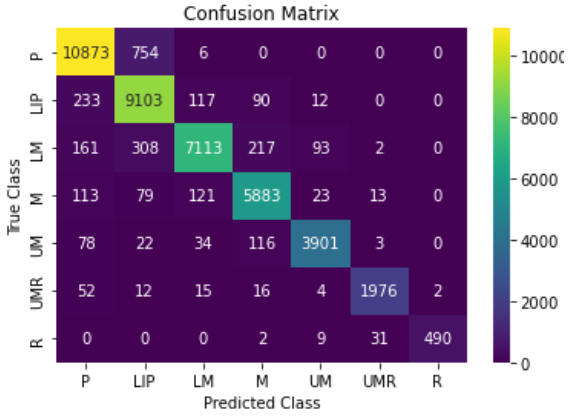

Figure 3. Confusion Matrix Report of the Model I

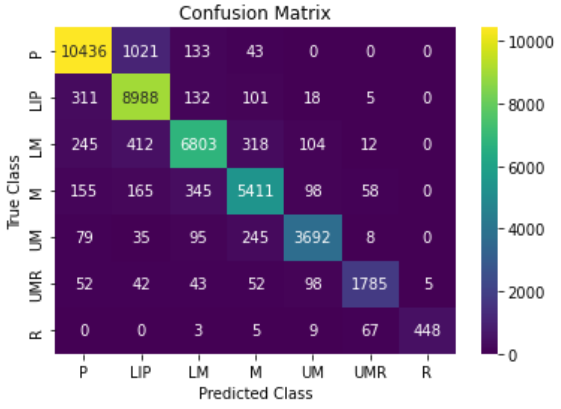

Figure 4. Confusion Matrix Report of the Model II

The predictive performance of Model I yields to 93% accuracy as shown in the Table VI and Table VII. The computation resulted to the precision of 0.93, recall of 0.94 and F1-score of 0.94 weighted mean respectively. Furthermore, Model II yields to 89% accuracy. The computation resulted in the classification report to a precision of 0.90, recall of 0.93 and , f1-score of 0.91 weighted mean respectively.

TABLE VI. CLASSIFICATION REPORT FOR MODEL I USING BAGGING ENSEMBLE

Income Class	*Precision*	*Recall*	*F1 - score*
Poor	0.93	0.94	0.94
Low-income but not poor	0.95	0.89	0.92
Lower Middle	0.90	0.96	0.93
Middle	0.94	0.93	0.94
Upper-Middle	0.94	0.97	0.95
Upper-middle but not rich	0.95	0.98	0.96
Rich	0.92	1.00	0.96
Weighted Mean	0.93	0.94	0.94
Accuracy			**0.93**

TABLE VII. CLASSIFICATION REPORT FOR MODEL II USING BOSTING ENSEMBLE

Income Class	*Precision*	*Recall*	*F1 - score*
Poor	0.90	0.93	0.91
Low-income but not poor	0.94	0.84	0.89
Lower Middle	0.86	0.90	0.88
Middle	0.87	0.88	0.87
Upper-Middle	0.89	0.92	0.90
Upper-middle but not rich	0.86	0.92	0.89
Rich	0.84	0.99	0.91
Weighted Mean	0.90	0.93	0.91
Accuracy			**0.89**

In this research, Model II(Boosting) yielded lower accuracy of 0.89 than of Model I(Bagging) with 0.93. Therefore, Bagging Naive Bayes gives the highest accuracy score of 93% while Boosting Naive Bayes yielded only 89%. According to Lima Vallantin, a model with an accuracy of 80% above is considered a good model [24]. It is evident that the method employed in this research and the final results obtained are completely acceptable in the scientific literature [25].

The findings of this study potentially support the local government unit in determining which households should receive priority for programs aimed at reducing poverty. The outcome can also make it easier for policymakers to identify relevant community-based policies and strategies for reducing poverty.

4. CONCLUSION AND RECOMMENDATIONS

In this study, the researchers presented a Naïve Bayes model for predicting Filipino income classes using selected FIES dataset provided by the Philippine Statistics Authority (PSA). Bagging and Boosting ensemble methods were used to produce the most optimal predictive model. The model was evaluated using various model evaluation metrics such as confusion matrix, accuracy, precision, recall, and F1-score. Results showed that the Naive Bayes algorithm demonstrated effectiveness, with the bagging ensemble yielding higher accuracy (93%), precision (93%), recall (94%), and F1-score (0.94%) compared to the boosting ensemble with (89%) accuracy. This indicates that the Naive Bayes model can serve as a practical alternative to other predictive models that utilize the FIES dataset, providing practical insights to determine income classes and identify factors strongly correlated with the target variable. Furthermore, additional parameters such as regions, family size, and others could potentially improve model accuracy. Exploring the use of other machine learning algorithms and ensembles could also enhance future work. Additionally, considering the increasing data of the Filipino Income and Expenditure Survey datasets provided by the Philippine Statistics Authority (PSA) every three years should be taken into account in future research. Overall, the findings of this study highlight the potential of using Naive Bayes models in predicting income classes in the Philippines and provide recommendations for further improvement and exploration in future research.

REFERENCES

[1] A. Greenspan, Poverty in the Philippines: the impact of family size. Asia Pac Pop Policy, PMID: 12317439, 1992.

[2] UNICEF, Child Poverty in the Philippines, United Nations Children's Fund (UNICEF) , 2015.

[3] PopCom, "PopCom," 2022. [Online]. Available: https://popcom.gov.ph/family-size-matters-average-filipino-family-spends-40-of-monthly-expenses-on-food/.

[4] PSA, "2021 Family Income and Expenditure Survey," Philippine Statistic Authority, 2021. [Online]. Available: https://psa.gov.ph/content/2021-family-income-and-expenditure-survey-adopts-computer-aided-personal-interview-capi-data.

[5] M. R. M. A. e. Jose Ramon G. Albert, "Poverty, the Middle Class, and Income Distribution," Philippine Institute for Development Studies, no. 2020-22, 2020.

[6] S. Liu, M. Zhu and Y. Yang, "A Bayesian Classifier Learning Algorithm Based on Optimization Model," Mathematical Problems in Engineering, vol. 2013, p. 9, 2013.

[7] H. Chen, S. Hu, R. Hua and X. Zhao, "Improved naive Bayes classification algorithm for traffic risk management," EURASIP Journal on Advances in Signal Processing, vol. 2021, no. 30, 2021.

[8] T. Mitchell, "GENERATIVE AND DISCRIMINATIVE CLASSIFIERS:NAIVE BAYES AND LOGISTIC REGRESSION," Machine Learning, 2020.

[9] D. Cousineau and S. Chartier, "Outliers detection and treatment: A review," International Journal of Psychological Research, 2010.

[10] S. I. Khan and A. S. M. L. Hoque, "SICE: an improved missing data imputation technique," Journal of Big Data, no. 37, 2020.

[11] A. Taha, B. Cosgrave and S. Mckeever, "Using Feature Selection with Machine Learning for Generation of Insurance Insights," MDPI, 2022.

[12] M. M. Usman, O. Owolabi and O. Owolabi, "Feature Selection: It Importance in Performance Prediction," IJESC, 2020.

[13] H. Osman, M. Ghafari and O. Nierstrasz, "The Impact of Feature Selection on Predicting the Number of Bugs," Arxiv, 2018.

[14] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," Frontiers : Integrative Bioinformatics, 2022.

[15] J. Li, K. Cheng, S. Wang and F. Morstatter, "Feature Selection: A Data Perspective," ACM Computing Surveys, vol. 50, no. 6, 2016.

[16] A. Clark, "The Machine Learning Audit - CRISP-DM Framework," ISACA Journal, 2018.

[17] Q. H. Nguyen, H.-B. Ly, L. S. Ho and et.al., "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil," Artificial Intelligence for Civil Engineering, vol. 2021, 2021.

[18] J. Roshan, "Optimal Ratio for Data Splitting," 2022. [Online]. Available: https://arxiv.org/pdf/2202.03326.pdf.

[19] H. Jafarzadeh, M. Mahdianpari, E. Gill and F. Mohammadimanesh, "Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral and PolSAR Data: A Comparative Evaluation," Remote Sensing, 2021.

[20] S. Raschka, J. Patterson and C. Nolet, "Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence," Information, vol. 11, no. 4, p. 193, 2020.

[21] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation," Journal of Machine Learning Technologies , 2011.

[22] Z. Vujovic, "Classification Model Evaluation Metrics," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, 2021.

[23] L.-j. Cai, S. Lv and K.-b. Shi, "Application of an Improved CHI Feature Selection Algorithm," Discrete Dynamics in Nature and Society, vol. 2021, 2021.

[24]L. Vallantin, "Accuracy to measure machine learning performance," 2018. [Online]. Available: https://medium.com/@limavallantin/why-you-should-not-trust-only-in-accuracy-to-measure-machine-learning-performance-a72cf00b4516. [Accessed June 2022].

[25] J. Talingdan, "Performance Comparison of Different Classification Algorithms for Household Poverty Classification,"