Performance Comparison of Different Classification Algorithms for Household Poverty Classification

Janelyn A. Talingdan
Information and Communications Technology Department
Abra State Institute of Sciences and Technology
Lagangilang, Abra, Philippines
janelyn.ambre@gmail.com

Abstract— **There should be a way to identify poor households as the potential receiver of poverty alleviation programs if the aim of an organization is to serve the deprive one. This problem is addressed by using machine learning algorithms particularly the classification algorithms. In this study, data mining was used to analyze the poverty data that was extracted from the Community Based Monitoring System (CBMS) Database of Lagangilang, Abra, Philippines. Different methods of data mining classification were evaluated and compared like Naïve Bayes, ID3, Decision Tree, Logistic Regression and KNearest Neighbor (KNN). In order to get the best predictive model, the different classification algorithms were evaluated based on the performance metrics; accuracy, precision, recall, F1 score and AUC. The rapid miner was the datamining tool used to process the different classification algorithms and to analyze the dataset. The study showed that Naïve Bayes classifier is an efficient algorithm for predicting households that are poor and non-poor because it outperformed all the four algorithms in all the performance metrics used in the study. Furthermore, the error rate of the said algorithm is significant because it is 0.0014 only. Therefore, choosing the best classification algorithm for data mining task will give a better result.**

Keywords-poverty, data mining, naïve bayes, performance metrics

I. INTRODUCTION

Poverty is a very complicated social phenomenon and it has become a large issue around the world. Poverty has many faces and it varies from place to place. Poverty is when people cannot afford some life's basic needs, such as food, shelter, clothing, basic education, primary health care and security [1]. When the aim of an organization is to serve the deprive one, then there should be a way to identify poor households as their potential receiver of poverty alleviation programs. This problem is addressed by using machine learning algorithms particularly the supervised learning to help in predicting households belong to above and below poverty line.

Machine learning enable to analyse the huge amount of data and extract the patterns which can be used for different applications. The evaluation of the historical data is termed in machine learning as training of the computational model [2]. Based on the experience collected from the historical records the future trends and upcoming events are predicted or approximated. Therefore, the data mining techniques supports the classification and prediction based on supervised learning concept for analysis of previous data. Basically the data mining techniques offered the analysis of the patterns of data and utilize them to develop classification, prediction and pattern recognition data models [3]. The supervised learning functions in two major modules training and testing. Training process evaluates the data pattern and during the testing the algorithm recognize the similar pattern data. The main advantage to use the supervised technique is their performance and accurate outcomes as compared to unsupervised approaches of learning [4]. Choosing the best model for a particular type of data mining is challenging that is why it is necessary to perform different techniques in order to choose the one that gives the best result. Therefore, the different algorithms should be evaluated based on their performance to delineate how good the predictive model are.

In this study, data mining will be used to analyze the poverty data that was extracted from the Community Based Monitoring System (CBMS) Database of Lagangilang, Abra, Philippines. The knowledge gained from mining the data of CBMS is very useful and it will support and assist the local government unit to determine households that are poor and non-poor. This will also help them in designing effective poverty reduction policies and programs. Different methods of data mining classification will be evaluated and compared like Naïve Bayes, ID3, Decision Tree, Logistic Regression and KNearest Neighbor. To get the best predictive model, the different classification algorithms will be evaluated based on the accuracy, precision, recall, F1 score and AUC. The rapid miner will be the datamining tool to use to process the different classification algorithms and to analyze the dataset.

II. LITERATURE REVIEW

This section reviews the different data mining techniques and how they were used for different studies for classification and prediction. The techniques that were

2160-1291/19/$31.00 ©2019 IEEE
DOI 10.1109/ICISE.2019.00010

reviewed were naïve bayes, decision tree, ID3, KNN and logistic regression.

The K-NN classifier and the Naive Bayes were used for crime prediction and classification in San Francisco. In the K-NN classifier, the uniform and inverse techniques were used while in the Naive Bayes, the gaussian, bernoulli, and multinomial techniques were tested. Validation and cross validation were used to test the result of each technique. The experimental results showed that higher classification accuracy ca be obtained using the multinomial Naive Bayes cross validation [5]

Another study used classification method to search alternative design to simulate energy use by a building prior to the erection of the building. The classifiers used were Naïve Bayes, Decision Tree, and k-Nearest Neighbor. The study showed that Decision Tree has the fastest classification time followed by Naïve Bayes and k-Nearest Neighbor. Based on precision, recall, f-measure, accuracy, and AUC, the performance of Naïve Bayes was the best. It outperformed Decision Tree and k-Nearest Neighbor on all parameters except precision [6].

The tree-structured method has been used in the study to classify poverty of a household and it gives a fairly accurate result. The benefit of the classification tree model is simpler and efficient. Instead of using a multiple regression equation they apply the classification tree incorporating survey weights for the prediction [7].

Another research implement and compare the results of several machine learning classification algorithms such as Random Forest, Support Vector Machine, and Logistic Regression to identify poverty for different block groups in the United States. Random Forests outperform Logistic Regression and Support Vector Machines. It is observed that random forest classifier predicts well and has better performance metrics. It is also observed that random forests are relatively faster when compared to other algorithms. Random Forests are scalable and can be used with huge dataset and dimensions/features, since random forests can easily be run in a distributed environment and there is no data dependency between different trees that are formed as part of the random forest ensemble [8].

Based on the related studies above different classification algorithms were used in various studies and have been tested to get their performance to know the best predictive model. In this study, the naïve bayes, ID3, decision tree, logistic regression and KNN were considered for testing and compared its accuracy, recall, precision and F1 score, and AUC to get the best predictive model to be used in the study.

The Naïve Bayes Classifier is a supervised machine learning technique used to take decision under the uncertain conditions as well as a statistical method for classification [9]. Decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions [10]. On the other hand, *KNearest Neighbor (KNN) Classifier* is one of the top ten algorithms used for the classification and regression. It is, also known as lazy learner or instance based, in that it stores all of the training samples and do not build a classifier until a new sample needed to be classified that makes predictions based on KNN labels assigned to test sample [11]. Furthermore, Logistic Regression is a logistic regression analysis to find the best-fitting and most parsimonious, yet reasonable, model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables. What distinguishes the logistic regression model from the linear regression model is that the outcome variable in logistic regression is categorical and most usually binary or dichotomous [8]. Lastly, ID3 algorithm is a classification algorithm based on Information Entropy, its basic idea is that all examples are mapped to different categories according to different values of the condition attribute set; its core is to determine the best classification attribute form condition attribute sets. The algorithm chooses information gain as attribute selection criteria; usually the attribute that has the highest information gain is selected as the splitting attribute of current node, in order to make information entropy that the divided subsets need smallest [12]

III. METHODS

The methodological approach of this study is composed of: the collection of data sets about poverty, identification of poverty indicators to predict poverty, formulation of the predictive model using the supervised machine learning algorithms, and the performance evaluation metrics applied to determine the best predictive model.

A. Data Collection

The poverty data of the local government unit of Lagangilang, Abra, Philippines using the Community Based Monitoring System which was conducted in 2017 was used as the dataset of the study. CBMS is an organized way of collecting information at the local level that includes the following categories: health, nutrition, housing, water and sanitation, education, income, employment and peace and order.

B. Data Preprocessing

The survey result of CBMS is rich in terms of variables and number of cases therefore it was preprocessed to eliminate attributes that are not important to the study to form a smaller data set. The summary of the 14 core

indicators was only selected as attributes from the 237 original attributes and all the 1780 instances were used as the final dataset of the study. This process also includes cleaning the data set that involves dealing with missing values, outliers and inconsistent values. Missing values were replaced by most used value in the data while the inconsistent values were fixed. The data was subdivided into two, 80% (1424 instances) of the data was used as the training data while the remaining 20% (356 instances) was used as the test data. The division of data is necessary in order to get the best predictive model in predicting the households that are poor and non-poor. The identified variables as poverty indicators are summarized in table 1.

Table 1: Identified variables as poverty indicators

Categories	Indicators	Values
1. Health and Nutrition	Malnutrition	Without member 0-5
		Without malnourished children
		With malnourished children
	Child death 0-5	Without member 0-5 and no child death
		Without child death
		With child death
	Death due to pregnancy	Not applicable
		Without death due to pregnancy related causes
		With death due to pregnancy related causes
2. Housing	Squatters	Formal settler
		Informal settler
	Makeshift Housing	Not living in makeshift housing
		Living in makeshift housing
3. Water and Sanitation	Safe water Supply	With access to safe water
		Without access to safe water
	Sanitary Toilet Facility	With access to sanitary toilet
		Without access to sanitary toilet
4. Basic Education	School Participation 6-17	No member 6-17
		All members attending school
		With member not in school
	Literacy	All members are literate
		With illiterate member
5. Income and livelihood	Income below Poverty threshold	Poor
		Non-poor
	Subsistent Poverty	Subsistently poor
		Subsistently non-poor
	Food shortage	Did not experience food shortage
		Experienced food shortage
	Unemployment	All members in the labor force are employed
		With unemployed members of the labor force
		No members of the labor force
6. Peace and order	Victims of crime	No victims of crime
		With victims of crime

C. *Data Mining*

Data mining software is one of a number of analytical tools for data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational database [13].

The data mining technique was used in this study to predict the households that are poor and non-poor. The Rapidminer data mining tool for classification was used to determine the performance of the five classification algorithms like Naïve Bayes, Decision Tree, KNN, Logistic Regression and ID3.

D. *Performance Evaluation*

To get the best predictive model in prediction, the different classification algorithms were evaluated based on the following:

A. Accuracy that measures the performance of each model that gives percentage of features that are predicted correctly among total number of features,

B. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.,

C. Recall that gives number of positive features classified correctly by the model,

D. F1 Score is a harmonic mean of precision and recall for balancing out both has been taken as a measure of performance.

E. AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

IV. RESULTS AND DISCUSSION

Table 2 shows the performance of the five classification algorithms used in this study such as accuracy, precision, recall, F1 score and AUC while Fig. 1 shows the comparative graph of the five algorithms based on their performance metrics.

The five algorithms Naïve Bayes, Decision Tree, KNN, Logistic Regression and ID3 were used in this study to predict the households that are poor and non-poor. These five algorithms were evaluated based on their performance metrics to get the bet predictive model. It can be seen in Table 2 that based on the accuracy of the five algorithms the Naïve Bayes out performs the other four algorithms with an accuracy rate of 98.86% while the Decision Tree has the lowest accuracy rate of 88.76%. In terms of precision, it is also noted that Naïve Bayes has the highest percentage of 99.47% and the lowest is still the Decision Tree with a rate of 70.15%. When it comes to recall, the Naïve Bayes and Decision Tree have the best performance rate of 100% and the KNN is the lowest among the five algorithms with 91.76% rate.

The Naïve Bayes and Decision Tree have the same recall performance rate which is 100%, therefore, the F1 measure was used to validate which classifier is really the best in terms of precision and recall. F1 measure is the harmonic mean of precision and recall. From this performance metric, it prevails that Naïve Bayes has the highest F1 measure of 99.73% while the Decision Tree got the lowest F1 measure rate of 82.46%. Even the Decision Tree has a perfect recall rate, it has a lowest F1 score because of very low precision rate of 70.15%.

To validate furtherly the performance of the five algorithms in order to determine the best classifier for the data set, the AUC measure was measured. The AUC score of Naïve Bayes which is 1.0 strengthen that it is the best algorithm for prediction because the score is near to 1 which means it has a good measure of separability.

Table 2: Classification Algorithms Performance

Classification Algorithm	Accuracy	Precision	Recall	F1 Score	AUC	Average
Naïve Bayes	99.86	99.47	100	99.73	1.00	80.012
Decision Tree	88.76	70.15	100	82.46	0.924	68.458
KNN	92.63	82.34	91.76	86.80	0.976	70.9012
Logistic Regression	89.04	70.75	99.73	82.77	0.947	68.6474
ID3	90.31	74.49	96.28	83.99	0.958	69.2056

Furthermore, Fig. 1 shows the comparative graph of the five algorithms and it shows that Naïve Bayes predicts well and has the best performance metrics compare to the four classification algorithms used in this study.

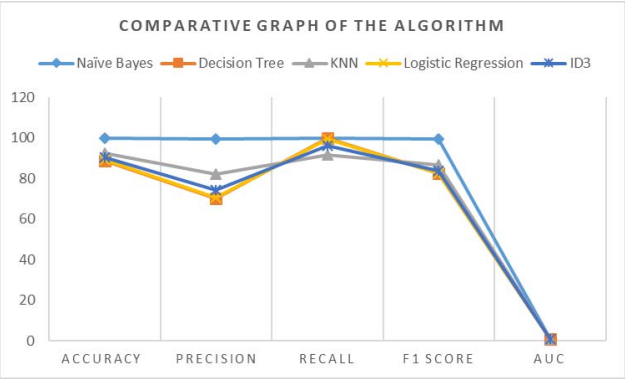

Figure 1. Comparative Graph of the Algorithms

After evaluating the different performance metrics of the algorithms, table 3 shows the confusion matrix of the Naïve Bayes which prevail as the best predictive model to be used in the study. Confusion matrix contains information about actual and predicted classification done by classification system. Performance of the system is commonly evaluated using data in matrix [14]. The total number of true positives is 1046 while the total number of false positive is 2. On the other hand, the total number of true negative is 376 and false negative is 0. The table shows also the error rate of the Naïve Bayes which is 0.0014.

Table 3. Confusion Matrix

	Poor	Non-Poor	Error rate
Poor	1046	0	0.0014
Non-Poor	2	376	

The result of the study could help the local government unit to identify the households that need to be prioritized for poverty alleviation programs. The result can also help the policy makers to easily identified appropriate poverty reduction policies and programs to be rendered in the community

V. CONCLUSION

Choosing the right algorithm for a particular type of data mining task is not easy. The best way is to perform a validation of the performances of the different algorithms to choose the best one that gives an appropriate result. This study performed a comparative analysis of the five classification algorithms; Naïve Bayes, Decision Tree, KNN, Logistic Regression and ID3 using the CBMS data set of Lagangilang, Abra, Philipppines. In order to get the best predictive model, the different classification algorithms were evaluated based on the performance metrics; accuracy, precision, recall, F1 score and AUC. The rapid miner was the datamining tool used to process the different classification algorithms and to analyze the dataset. It can be concluded that Naïve Bayes classifier is an efficient algorithm for predicting households that are poor and non-poor because it outperformed all the four algorithms in all the performance metrics used in this study. Furthermore, the error rate of the said algorithm is significant because it is 0.0014 only. Therefore, choosing the best classification algorithm for data mining task will give a better result.

ACKNOWLEDGMENT

The researcher wishes to express her genuine appreciation to all those who are, in one way or another, instrumental to the successful completion of this work.

Dr. Gregorio T. Turqueza Jr., President II of the Abra State Institute of Sciences and Technology, for his consistent support to the endeavors of the ASIST faculty researchers.

Dr. Noel B. Begnalen, Vice-President for Academic Affairs, for his unconditional provision for research funding and their encouragement.

Dr. Pablo B. Bose Jr., Director for Research and Development, for his unselfish dedication to research undertakings.

Dr. Mary Joan T. Guzman, Dean of the College of Arts and Sciences, for her constant sisterly and motherly counsel.

To all IT Faculty, she shared laughter and productive thoughts.

REFERENCES

[1] Kamanou, G., Morduch, J., Isidero, D. P., Gibson, J., Ivo, H., & Ward, M. (2005). Handbook on poverty statistics: Concepts, methods and policy use. The United Nations Statistics Division. Retrieved from http://unstats. un. Retrieved from org/unsd/methods/poverty/pdf/UN_Book% 20FINAL, 2, 030.

[2] El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In *Machine Learning in Radiation Oncology* (pp. 3-11). Springer, Cham.

[3] Chouhan, S. S., & Khatri, R. (2016). Data Mining based Technique for Natural Event Prediction and Disaster Management. International Journal of Computer Applications, 139(14).

[4] Chouhan, S. S., & Khatri, R. (2016). Data Mining based Technique for Natural Event Prediction and Disaster Management. International Journal of Computer Applications, 139(14).

[5] Abdulrahman, N., & Abedalkhader, W. KNNCLASSIFIER AND NAÏVE BAYSE CLASSIFIER FOR CRIME PREDICTION IN SAN FRANCISCO CONTEXT.

[6] Jeyarani, D. S., Anushya, G., & Pethalakshmi, A. (2013). A comparative study of decision tree and Naive Bayesian classifiers on medical datasets. *age*, *30*, 31-40.

[7] Bilton, P., Jones, G., Ganesh, S., & Haslett, S. (2017). Classification trees for poverty mapping. *Computational Statistics & Data Analysis*, *115*, 53-66.

[8] Korivi, K. (2016). Identifying poverty-driven need by augmenting census and community survey data.

[9] Tolan, G. M., & Soliman, O. S. (2015). An Experimental Study of Classification Algorithms for Terrorism Prediction. *International Journal of Knowledge Engineering-IACSIT*, *1*(2), 107-112.

[10] Ashari, A., Paryudi, I., & Tjoa, A. M. (2013). Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *4*(11).

[11] Gohar, F., Butt, W. H., & Qamar, U. (2014). Terrorist Group Prediction Using Data Classification. In *Work. MultiRelational Data Min. MRDM2003* (Vol. 10, pp. 199-208).

[12] Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Predicting students' performance using ID3 and C4. 5 classification algorithms. *arXiv preprint arXiv:1310.2071*.

[13] Pandey, A. (2014). Study and Analysis of K-Means Clustering Algorithm Using Rapidminer. International Journal of Engineering Research and Applications, 4(12), 60-64.)

[14] COE, J. (2012). Performance comparison of Naïve Bayes and J48 classification algorithms. *International Journal of Applied Engineering Research*, *7*(11), 2012.