Comparations of Supervised Machine Learning Techniques in Predicting the Classification of the Household's Welfare Status

Perbandingan Teknik Pembelajaran Mesin Terawasi dalam Memprediksi Klasifikasi Status Kesejahteraan Rumah Tangga

Nofriani

BPS-Statistics of Bengkulu Province
Jl. Adam Malik, Km. 8, Bengkulu City

nofriani@bps.go.id

Diterima : 11 Januari 2019 || Revisi : 20 April 2019 || Disetujui: 22 April 2019

Abstract - Poverty has been a major problem for most countries around the world, including Indonesia. One approach to eradicate poverty is through equitable distribution of social assistance for target households based on Integrated Database of social assistance. This study has compared several well-known supervised machine learning techniques, namely: Naïve Bayes Classifier, Support Vector Machines, K-Nearest Neighbor Classification, C4.5 Algorithm, and Random Forest Algorithm to predict household welfare status classification by using an Integrated Database as a study case. The main objective of this study was to choose the best-supervised machine learning approach in predicting the classification of household's welfare status based on attributes in the Integrated Database. The results showed that the Random Forest Algorithm was the best.

Keywords: classification, data training and testing, k-fold cross-validation, integrated database, supervised machine learning

Abstrak - *Kemiskinan merupakan permasalahan besar bagi banyak negara di dunia, termasuk Indonesia. Salah satu pendekatan untuk memberantas kemiskinan adalah melalui distribusi merata bantuan sosial untuk rumah tangga sasaran dengan berbasis pada Data Terpadu. Penelitian ini membandingkan antara beberapa teknik Pembelajaran Mesin Terawasi yang umum digunakan, yakni: Naïve Bayes Classifier, Support Vector Machines, K-Nearest Neighbor Classification, C4.5 Algorithm dan Random Forest Algorithm untuk memprediksi status kesejahteraan rumah tangga dengan menggunakan Basis Data Terpadu sebagai studi kasus. Tujuan utama penelitian ini adalah untuk memilih pendekatan supervised machine learning yang paling baik dalam memprediksi klasifikasi status kesejahteraan rumah tangga berdasarkan atribut dalam Basis Data Terpadu. Hasil penelitian menunjukkan bahwa Random Forest Algorithm adalah yang terbaik.*

Kata kunci*: basis data terpadu, klasifikasi, pembelajaran mesin terawas, pengujian data, validasi k-fold cross*

INTRODUCTION

Poverty has been a major problem for most countries around the world, especially for the last couple of decades. In Indonesia, it is one of the primary issues for the Government to solve. One of the approaches the Government has attempted to eradicate poverty in Indonesia was the distribution of social assistance across the archipelago for the target households those were categorized as worth receiving social assistance according to the integrated database. The data from all regions in Indonesia, including from Bengkulu as one of its less developed provinces, is collected and updated every certain year. It includes various variables that contribute to classifying the household's welfare status.

A large number of works have performed comparative studies of supervised machine learning methods using real or dummy data. Hastuti (2012) studied the prediction of inactive university students using Decision Tree C4.5 algorithm on 3,861 data sets with 21 attributes. The research showed that the algorithm gives 95.29% accuracy. Another research conducted by Defiyanti (2014), used C4.5 and IDE3 methods in classifying email's spam with a variety of amounts of attributes and data set. It found that C4.5 gave the highest accuracy of 72.38% with 52 attributes and IDE3 gave 73.20% with 58 attributes.

Despite the fact that Integrated Database plays an important role in poverty eradication effort, only a few types of research have used Integrated Database as a case study in classification algorithm analysis. Research by

DOI:10.30818/jpkm.2019.2040105

Karyadiputra (2016) obtained classification accuracy of 85.80% and AUC (Area Under the Curve) value of 0.930 in predicting household's welfare status using Naïve Bayes Classifier on Integrated Database of 2011 with 16 attributes. Another research by Iskandar (2013) compared two classification algorithms: C4.5 and Naïve Bayes in predicting household's welfare status on Integrated Database of 2015 with 13,928 data sets. It concluded that the C4.5 algorithm had a higher accuracy by 3% due to its accuracy of 64%, while the Naïve Bayes algorithm obtained an accuracy of only 61%.

Integrated Database is also used in this research as a case study in predicting household's welfare status. However, unlike previous related research, this research uses and compares more than two supervised learning approaches. This research aims to choose the best practical supervised learning approach in predicting the household's welfare status using contributing attributes of the Integrated Database.

The detailed objectives of the current research are to prove whether or not the supervised learning approach can be used to predict household's welfare status based on the criteria provided in the Integrated Database, to choose best practical supervised machine learning approach in predicting the classification of household's welfare status based on analysis on classification accuracy and other evaluation methods for classification algorithms, and to provide the best data classifier model to predict household's welfare status.

The current research scope only included the well-known and frequently used classification algorithms of supervised machine learning techniques, such as The Naïve Bayes Classifier, the Support Vector Machines, the K-Nearest Neighbor Classification, the C4.5 Algorithm, and the Random Forest Algorithm. It only covered the comparative analysis on classification techniques of supervised machine learning, without attempting to judge or evaluate the data reliability and validity of the Integrated Database of 2015. Furthermore, the scale of the Integrated Database is also limited to the provincial level, namely Bengkulu, one of the less developed provinces in Indonesia.

METHODOLOGY

For better results of the analysis, this paper provides a brief literature review of the Integrated Database and supervised machine learning.

Integrated Database

Integrated Database is a fundamental guideline book for the Indonesian Government for better distribution of social assistance. The book that is released by TNP2K (National Team for Acceleration of Poverty Eradication) is used to improve the quality of the target determination process of social assistance programs. It has over 50 attributes that determine each household's poverty level; such as education, occupation, the status of residence, condition of residence, number of household members, etc. (TNP2K, 2018). Its main source was the results of updates those released by Statistics of Indonesia (BPS) in 2015 and the collaboration with the Social Ministry of the Republic of Indonesia (Iskandar, 2013). By using Integrated Database, the total amount and detailed information of target households for social assistance program can be properly analyzed from the very beginning of the program planning, so then it surely helps to reduce the error of target determination.

Supervised Machine Learning

Machine learning uses computers to simulate human learning and allows them to identify and acquire knowledge from the real world and improve performance on some tasks based on this new knowledge (Portugal et al, 2018). Mitchell, et al (2013) define machine learning as: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".

Supervised machine learning is when one has input variables (x) and an output variable (y) using an algorithm to learn the mapping function from input to the output, with the formula $Y = f(X)$. In it, all data is labeled beforehand and the algorithms learn to predict the output from the input data (Brownlee, 2018). This is intended to estimate the mapping function as best as possible so that when someone later has a new input data (x), he can predict the output variable (y) of that data.

The supervised learning matter can be further grouped into the regression and the classification. A regression matter is when the output variable is a real value, like dollars or weight. While classification matter is when the output variable is a category, like 'red or blue' and 'yes or no' (Brownlee, 2018). Figure 1 shows a diagram of the model building procedure for data classification.

Supervised learning algorithm (such as classification) is more preferred than unsupervised learning algorithm (such as clustering), because its prior knowledge of the data records class labels makes the feature/attribute selection easier, and leads to better prediction/classification accuracy (Anyanwu, 2009).

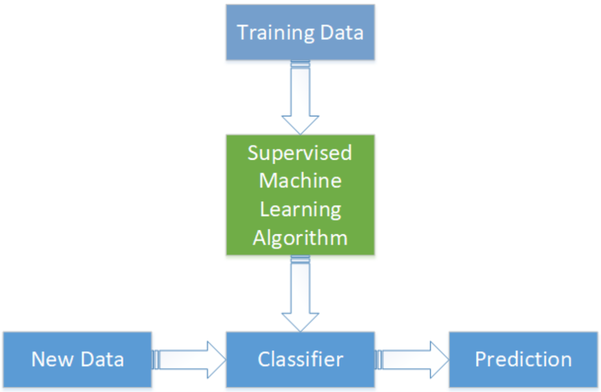


Figure 1 Diagram of General Model Building Procedure for Data Classification

Naïve Bayes Classifier

A Naïve Bayes Classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naïve) independence assumption (Murphy, 2006). Its advantage is that it only requires a small amount of training data to estimate the parameters necessary for classification (Kaur, 2014). Its disadvantage is that it does not involve morphological relation among the features or terms. It is used in personal email sorting, document categorization, email spam detection, and sentiment detection (Kini, 2015).

Support Vector Machines

Support Vector Machines (SVM) is a supervised machine learning technique which is based on statistical theory (Janardhanan, 2015). A set of the training set is marked as belonging to one or two categories in it. Its training algorithm builds a model that assigns new data set to one category or making it a non-probabilistic binary linear classifier.

It is widely used in medical imaging, image interpolation, medical classification tasks, financial analysis, neural networks, pattern recognition, and page ranking algorithm. Its disadvantage is that it gives poor performance when the number of features (variable x) is bigger than the number of samples (YouTube, 2018).

K-Nearest Neighbor Classification

K-Nearest Neighbor (KNN) has been generally used in GEOBIA workflows (Luque, 2013) due to its simplicity and flexibility (Li, 2016). Understanding KNN Classification is quite simple, examples are classified based on the class of their nearest neighbors. It is a type of instance-based/lazy learning where the function is only approximated locally, and all computation is delayed until classification. In the KNN rule, a test set is assigned the class most frequently represented among the k nearest training set. If two or more such classes exist, then the test set is assigned the class with a minimum average distance to it (Kataria, 2013).

KNN Classification can be calculated mostly by calculating Euclidian distance. Although other measures are also available, through Euclidian distance one has splendid intermingle of ease, efficiency, and productivity (Podgorelec, 2002). It can be used in text mining or text categorization, climate forecasting, estimating soil-water parameters, stock market forecasting, medical disease prediction, etc. Its advantages are robust to noisy training data, effective for large training data set, and learns complex models easily, while its disadvantage is that it is difficult to determine the value of parameter k in high-dimensional data.

C4.5 Decision Tree

A decision tree is a type of supervised machine learning algorithm that is mostly used in classification matters. It is a flowchart-like structure which each internal node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf or terminal node holds the class label. It can be used on numerical or categorical data. It is simple to understand, interpret and visualize. Its disadvantage is possible overfitting, in which decision trees create over complex trees that do not generalize the data well. It can be unstable because of the small variations in the data may result in a completely different tree.

Decision trees make use of the IDE3 (Iterative Dichotomiser 3), C4.5 and CART algorithms. In the IDE3, data is sorted at every node during the tree building phase, in order to select the best splitting single attribute. It does not give accurate results when there is too much noise or details in the training data set, thus an intensive pre-processing of data is carried out before building a decision tree model with the IDE3. C4.5 is an extension of the IDE3 algorithm. Pruning takes place in C4.5 by replacing the internal node with a leaf node, thereby reducing the error rate (Podgorelec, 2002). Unlike the IDE3, the C4.5 accepts both continuous and categorical attributes in building the decision tree. It has an enhanced method of tree pruning that reduces misclassification errors, due to too much noise or details in the training data set. Like in the IDE3, the data is sorted at every node of the tree, in order to determine the best splitting attribute. It uses the gain ratio impurity method to evaluate the splitting attribute (Quinlan, 1986). This research chooses the C4.5 algorithm over the IDE3 algorithm because research by Chauhan (2014) found that the C4.5 algorithm outperforms the IDE3 algorithm.

Random Forest Algorithm

Random Forest (RF) Algorithm is one of the most popular and most powerful supervised machine learning algorithms that are capable of performing both regression and classification tasks. This algorithm creates a forest with a number of decision trees. In general, the more tree there are in the forest, the more robust the prediction will be, and thus the higher the accuracy obtained. To classify a new object based on attributes, each tree gives a classification vote. The forest chooses the classification having the most votes of all the trees in the forest (YouTube, 2018). Since the RF classifier was proposed, it has been improved continuously in the field of remote sensing image information extraction, where it has been shown to be a robust classifier (Chan, 2008).

Its advantages are: it handles the missing values and maintains accuracy for a large amount of data, and also handles large data set with higher dimensionality without overfitting the model. However, overfitting might happen if there is too much noise in the data. It is commonly used in banking to find loyal customers out of fraud ones; in medical world to find correct components to validate medicine, or to analyze a patient's disease based on their medical records; in stock market to analyze the stock market behavior as well as expected loss or profit; in e-commerce to help customers by recommending products; and in computer vision such as Microsoft, besides that, it is also used for image classification in Xbox Console.

Data Source

The current research used secondary data source from Indonesian Integrated Database Updates 2015 for Bengkulu. The data set has only one variable as the classification label, namely Household's Welfare Status, and over 50 variables as attribute labels which are all nominal data. The detailed information on the class labels and attributes used in this research is depicted in Table 1.

Table 1 Class and Relevant Attributes of the Integrated Database

Variable	Explanation
Class: Welfare Status	1: Household in the lowest 10 percent of welfare status. Included in decile 1 (Very poor) 2: Household in the lowest 11to 20 percent of welfare status. Included in decile 2 (Poor) 3: Household in the lowest 21 to 30 percent of welfare status. Included in decile 3 (Near poor) 4: Household in the lowest 31 to 40 percent of welfare status. In decile 4 (Vulnerable)
X_1	Highest level of education
X_2	Occupation or business field
X_3	Status on primary occupation
X_4	Status of residential building
X_5	Status of residential land
X_6	Residential building's floors area
X_7	Residential building's floors type
X_8	Residential building's walls type
X_9	Residential building's walls condition
X_{10}	Residential building's roof type
X_{11}	Residential building's roof condition
X_{12}	Number of bedroom in residential building
X_{13}	Source of drinking water
X_{14}	Way to access drinking water
X_{15}	Primary lighting source
X_{16}	Type of installed electrical power
X_{17}	Cooking fuel/utensil
X_{18}	Type of defecation facility
X_{19}	Toilet type
X_{20}	Type of final fecal disposal facility
X_{21}	Ownership status of gas cylinders with a capacity of 5.5 kg or above
X_{22}	Ownership status of the refrigerator
X_{23}	Ownership status of the air conditioner
X_{24}	Ownership status of water heater
X_{25}	Ownership status of the house phone
X_{26}	Ownership status of television
X_{27}	Ownership status of computer or laptop
X_{28}	Ownership status of bicycle
X_{29}	Ownership status of the motorcycle
X_{30}	Ownership status of the car
X_{31}	Ownership status of the boat
X_{32}	Ownership of outboard motor
X_{33}	Ownership of motorboat
X_{34}	Ownership of ship
X_{35}	Number of the owned active phone number
X_{36}	Number of owned LCD TV
X_{37}	Ownership status of land asset
X_{38}	The total area of owned land asset
X_{39}	Ownership status of the house beside the residential building
X_{40}	Number of owned cow
X_{41}	Number of owned buffalo
X_{42}	Number of owned horse
X_{43}	Number of owned pig
X_{44}	Number of owned goat
X_{45}	Number of a household member

Data Pre-processing

Before the data is ready to be trained and tested, pre-processing is needed to make the classifier work better. Unused variables that are not relevant to the research scope are removed from the file to create better classification accuracy for all of the classification methods.

The current research uses an open source tool, namely Weka Application Version 3.8.2 to run the algorithm and better evaluate the results of each algorithm. Weka is a free open-source software containing a collection of machine learning algorithms for data mining tasks developed by Waikato University of New Zealand (Waikato, 2018) Because it uses the ARFF file format as the source file to do the data processing (Noviyanto, 2015), the original database file is first converted to CSV file to generate the ARFF file.

Table 2 The proportion of Training Set and Testing Set

	Type of set	
Testing Set	**Training Set**	**Total Data Set**
1 Fold	9 Folds Combined	
2,388	21,484	23,872

The data contains 23,872 fields. Each field already has a class variable. Therefore, this research used both training set and testing set out of the data that has already filled with classification labels. The training set and testing set are divided using k-fold cross-validation method. It is a resampling procedure used to evaluate machine learning models on a limited data sample such as the data used in this research. This approach involves randomly dividing the set of observations into *k* groups or folds, of approximately equal size. One fold is treated as testing (validation) set, while the rest *k-1* folds act as the training set. Typically, one performs k-fold cross validation using *k* = 5 or *k* = 10 as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance. (James, et al) For this reason, this research chooses *k* = 10.

The total 23,872 fields of data set are split into 10 sets. One by one, a set is selected as the testing set and 9 remaining ones are combined as the training set. The training-testing process is then repeated ten times using a different testing set (different fold). Table 2 shows the detailed proportion of training set and testing set out of the original data set obtained using k-fold cross-validation.

Data Training and Testing

The k-fold cross validation is used in evaluating the performance of each of the five chosen classification algorithms mentioned in the Introduction Section, which includes training and testing processes. Each process of data training results in a classifier model with respective classification accuracy. Each model will be applied to classify future data set. Figure 2 shows a flowchart of the overall methodology used in this research.

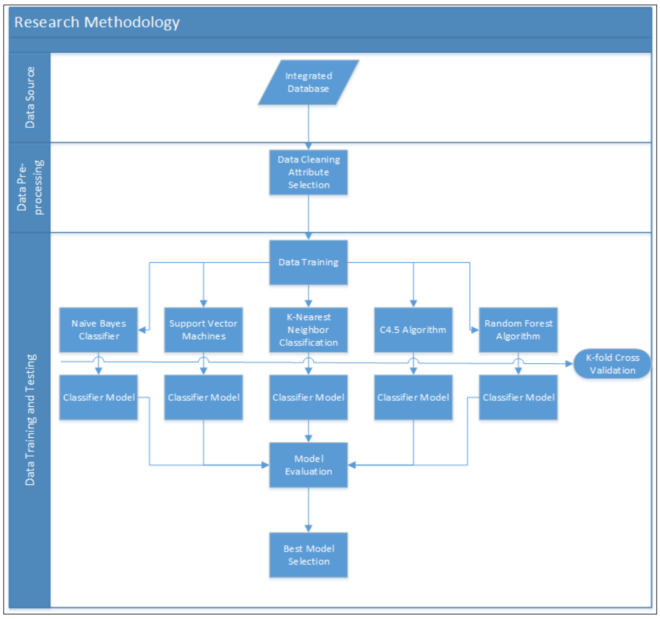


Figure 2 Flowchart of Research Methodology

RESULT AND ANALYSIS

Each classifier model is evaluated to choose the best one by comparing several indicators; such as confusion matrix, classification accuracy, precision and recall, and AUC.

Confusion Matrix

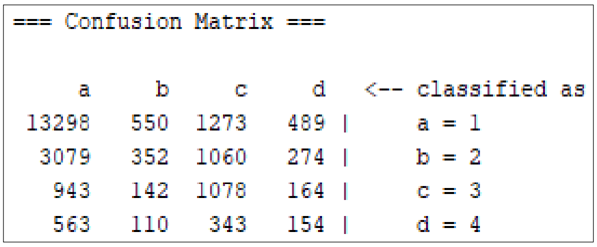
In this research, we investigate the use of the confusion matrix for attribute selection. A confusion matrix of size *n* x *n* associated with a classifier shows the predicted and actual classification, where n is the number of different classes (Visa, 2011). Table 3 shows a confusion matrix for *n* = 2, whose entries have the following meanings:

1. a is the number of correct negative predictions;
2. b is the number of incorrect positive predictions;
3. c is the number of incorrect negative predictions;
4. d is the number of correct positive predictions.

Table 3 The Confusion Matrix for Two-Class Classification Problem (*n*=2)

	Predicted Negative	**Predicted Positive**
Actual Negative	a	b
Actual Positive	c	d

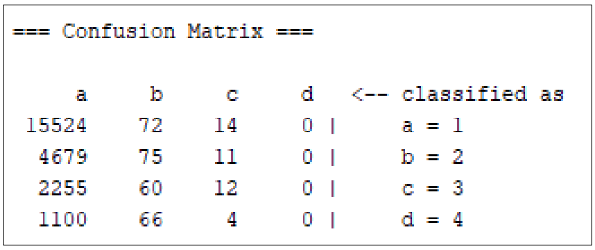
Figure 3 to Figure 7 show the confusion matrices for 1) Naïve Bayes Classifier, 2) Support Vector Machines, 3) K-Nearest Neighbor Classification, 4) C4.5 Algorithm, and 5) Random Forest Algorithm, respectively. These confusion matrices are generated in Weka application for each classification algorithm.



```
=== Confusion Matrix ===

     a     b     c     d   <-- classified as
 13298   550  1273   489 |     a = 1
  3079   352  1060   274 |     b = 2
   943   142  1078   164 |     c = 3
   563   110   343   154 |     d = 4
```

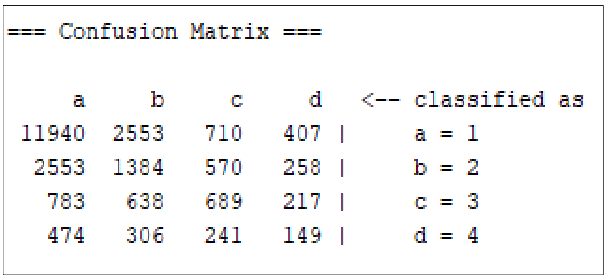
Figure 3 Confusion Matrix for Naïve Bayes Classifier



```
=== Confusion Matrix ===

     a     b     c     d   <-- classified as
 15524    72    14     0 |     a = 1
  4679    75    11     0 |     b = 2
  2255    60    12     0 |     c = 3
  1100    66     4     0 |     d = 4
```

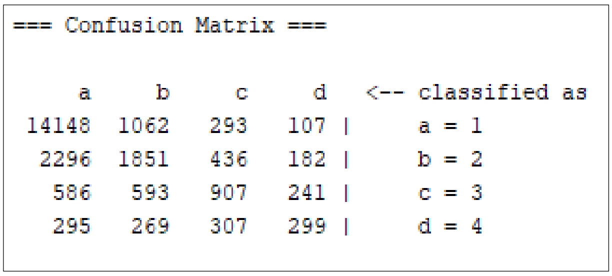
Figure 4 Confusion Matrix for Support Vector Machines



```
=== Confusion Matrix ===

     a     b     c     d   <-- classified as
 11940  2553   710   407 |     a = 1
  2553  1384   570   258 |     b = 2
   783   638   689   217 |     c = 3
   474   306   241   149 |     d = 4
```

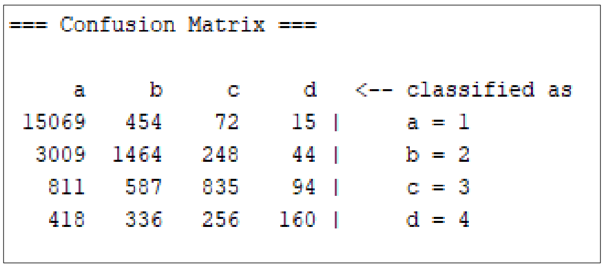
Figure 5 Confusion Matrix for K-Nearest Neighbor Classification



```
=== Confusion Matrix ===

     a     b     c     d   <-- classified as
 14148  1062   293   107 |     a = 1
  2296  1851   436   182 |     b = 2
   586   593   907   241 |     c = 3
   295   269   307   299 |     d = 4
```

Figure 6 Confusion Matrix for C4.5 Algorithm



```
=== Confusion Matrix ===

     a     b     c     d   <-- classified as
 15069   454    72    15 |     a = 1
  3009  1464   248    44 |     b = 2
   811   587   835    94 |     c = 3
   418   336   256   160 |     d = 4
```

Figure 7 Confusion Matrix for Random Forest Algorithm

From the figures above, it is understood that there are only three algorithms that have the highest number of correct predictions, i.e. SVM, C4.5, and Random Forest Algorithms. Out of 23,872 sets of attributes; SVM gives 15,611 correct predictions, C4.5 Algorithm gives 17,205 correct predictions, whilst the Random Forest Algorithm gives 17,528 correct predictions. This indicates that these three algorithms are more likely to be reliable in predicting the household's welfare status from the Integrated Database given the highest rates of correct predictions from the three algorithms. However, SVM is relatively less reliable, for only giving a large number of correct predictions for class variable = 1, while giving less or no correct predictions for the other three class variables.

Classification Accuracy

A further way to evaluate and compare classifiers is to calculate the prediction accuracy and classification error. Both can be obtained from a confusion matrix depicted in Table 3, as follows:

$$Accuracy = \frac{a+d}{a+b+c+d} \qquad (1)$$

$$Error = \frac{b+c}{a+b+c+d} \qquad (2)$$

Table 4 shows the summary of prediction accuracies and classification errors obtained from the five used algorithms.

Table 4 Prediction Accuracy and Classification Errors

Algorithm	Prediction Accuracy	Classification Error
Naïve Bayes Classifier	62.34 %	37.66 %
Support Vector Machines	65.40 %	34.60 %
KNN Classification	59.32 %	40.68 %
C4.5 Algorithm	72.07 %	27.93 %
Random Forest Algorithm	73.42 %	26.58 %

The table shows that there are two algorithms that give relatively higher prediction accuracy, i.e. C4.5 and Random Forest Algorithms. They give quite a high accuracy i.e. above 70%, which means there is a probability of 70 percent correct when they are used to classify the household's welfare status. This indicates that these two algorithms can relatively better predict household's welfare status in Integrated Database.

Classification Precision and Recall

A more detailed way to evaluate a classifier is to calculate the precision and recall of each class in classification. Precision is a measure of accuracy provided that a specific class has been predicted. It attempts to determine the proportion of positive identification that was actually correct. It is mathematically defined by:

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

where TP and FP are the numbers of true positive and false positive predictions for the considered class (Janardhanan, 2015). A model that produces no false positives has a precision of 1.0.

The recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. It is commonly called as sensitivity and corresponds to the true positive rate. It attempts to determine the proportion of actual positives that were identified correctly. It is mathematically defined by:

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

TP and FN are the numbers of true positive and false negative predictions for the considered class. TP + FN are the total numbers of test examples of the considered class (Janardhanan, 2015). A model that produces no false negatives has a recall of 1.0.

The way to interpret precision and recall is as follows:

1. If a model gives the precision of 0.75 when predicting a household's welfare status as near poor (class = 3), it is 75% correct of the time.
2. And if a model has a recall of 0.80 for a status of near poor (class = 3), it correctly identifies 80% of all statuses of near poor.

Precision and recall are often in tension. Improving precision typically reduces recall and vice versa (Google Developers, 2018). Nevertheless, to better evaluate the effectiveness of a classification model, both precision and recall must be examined. Because the higher the precision and recall, the better a classification model is. Table 5 shows the complete list of precision and recall for each class in the classifications using each algorithm.

Figure 8 shows the graph of the weighted average of precision and recall given by the five algorithms. From Table 5 and Figure 8, it is perceived that Precision and Recall are produced best out of two algorithms; C4.5 and Random Forest Algorithms. But both scores of Random Forest Algorithm is a little higher than of C4.5. This indicates that Random Forest is highly probable to be correct in predicting each class of household's welfare status, whilst C4.5 is a little less probable.

Table 5 Precision and Recall

Algorithm	Class	Precision	Recall
Naïve Bayes Classifier	1	0.744	0.852
	2	0.305	0.074
	3	0.287	0.463
	4	0.041	0.132
	Weighted Average	0.582	0.623
Support Vector Machines	1	0.659	0.994
	2	0.275	0.016
	3	0.293	0.005
	4	0.000	0.000
	Weighted Average	0.514	0.449
KNN Classification	1	0.758	0.765
	2	0.284	0.290
	3	0.312	0.296
	4	0.145	0.127
	Weighted Average	0.590	0.593
C4.5 Algorithm	1	0.817	0.906
	2	0.490	0.388
	3	0.467	0.390
	4	0.361	0.256
	Weighted Average	0.695	0.721
Random Forest Algorithm	1	0.780	0.965
	2	0.515	0.307
	3	0.592	0.359
	4	0.511	0.137
	Weighted Average	0.696	0.734

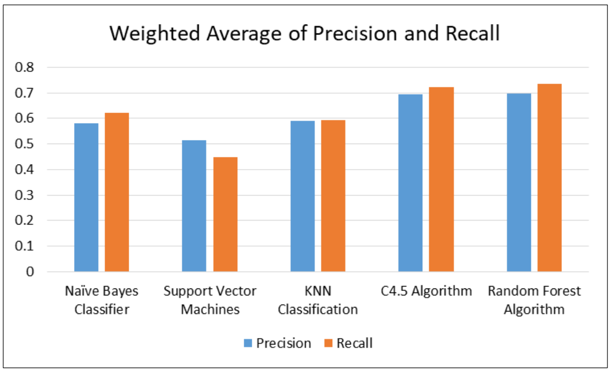


Figure 8 Weighted Average Score of Precision and Recall

AUC

Machine learning also uses the concept of ROC Curve (Receiver Operating Characteristic Curve) to evaluate a classification model. A ROC Curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots TP Rate (True Positive Rate) and FP Rate (False Positive Rate) at different classification thresholds. In the ROC Curve, the main goal is to have this curve more to the upper left corner, which is (0,1), (YouTube, 2013).

AUC, which stands for Area under the ROC Curve, measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). Figure 9 shows the AUC under the ROC curve.

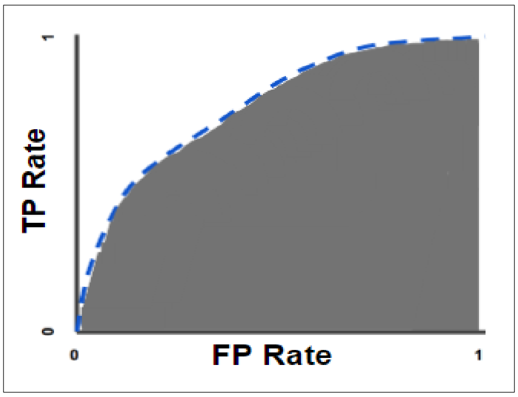


Figure 9 AUC (Area Under the ROC Curve)

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the classification model will rank a randomly chosen positive example higher than a randomly chosen negative example.

AUC ranges from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; whilst one whose predictions are 100% correct has an AUC of 1.0. AUC is preferable to use because it measures how well predictions are ranked, rather than their absolute values (Google Developers, 2018).

A rough guide for classifying the accuracy of a diagnostic test using AUC is the traditional system (Gorunescu, 2011), presented below:

1. 0.90 – 1.00 : excellent classification;
2. 0.80 – 0.90 : good classification;
3. 0.70 – 0.80 : fair classification;
4. 0.60 – 0.70 : poor classification;
5. 0.50 – 0.60 : failure.

Table 6 shows the summary of AUC value and interpretation obtained from the five algorithms. The table shows that Naïve Bayes Classifier gives three fair and one failed classification; SVM and KNN give either poor or failed classifications. C4.5 gives good classification for one class only, and poor for the other three. Meanwhile, Random Forest gives only good classifications for three classes, and fair for only one class. Therefore, in this part of model evaluation, Random Forest performs best for giving better AUC scores.

Table 6 AUC Interpretation of Each Algorithm

Algorithm	Threshold	AUC	Classification Interpretation
Naïve Bayes Classifier	1	0.722	Fair
	2	0.558	Failure
	3	0.707	Fair
	4	0.700	Fair
Support Vector Machines	1	0.593	Failure
	2	0.502	Failure
	3	0.617	Poor
	4	0.542	Failure
KNN Classification	1	0.652	Poor
	2	0.552	Failure
	3	0.612	Poor
	4	0.544	Failure
C4.5 Algorithm	1	0.806	Good
	2	0.670	Poor
	3	0.683	Poor
	4	0.618	Poor
Random Forest Algorithm	1	0.888	Good
	2	0.783	Fair
	3	0.857	Good
	4	0.871	Good

CONCLUSION

Five well-known supervised machine learning techniques have been analyzed to predict the classification of household’s welfare status in Integrated Database of 2015 for Bengkulu Province, Indonesia. The data training and testing use 23,872 fields of data set, divided into 2,388 fields of the training set and 21,484 testing set. The selection of the training set is performed ten times using k-fold cross-validation with the value of $k = 10$.

The current research proved that supervised machine learning techniques could be used for predicting household’s welfare status in Integrated Database using 45 attributes, with a variety of performance indicators. Three out of five algorithms give poor performances, i.e. Support Vector Machines (SVM), Naïve Bayes Classifier, and K-Nearest Neighbor Classification, which SVM relatively performed better than the other two. The remaining two give good performances, i.e. C4.5 and Random Forest Algorithms; which Random Forest gives quite better performance in all of the evaluation methods. In other words, Random Forest is the best practical choice of supervised machine learning technique in predicting household’s welfare status in Integrated Database of 2015. Additionally, the task of this research scope is far from perfect, future work is to build a system that is better at providing predictions.

ACKNOWLEDGEMENT

I would like to express my gratitude to Chief Statistician of Bengkulu Province, Mrs. Dyah Anugrah Kuswardani, M.A. who granted the permission to explore Integrated Database. I thank Mr. Dian Putra Nugraha, S.ST., for sharing his pearls of wisdom during the course of this research. I also thank Mr. Novrian Pratama, S.ST., M.Si. and Ms. Rahmah Ulfah, S.ST., M.Sc. who greatly provided insights and expertise that surely assisted this research.

REFERENCES

Anyanwu, M. N., & Shiva, S. G. (2009) Comparative Analysis of Serial Decision Tree Classification Algorithms. *International Journal of Computer Science and Security (IJCSS)*, 3(3), 230-240.

Brownlee, J. (2016). Supervised and unsupervised machine learning algorithms. *Machine Learning Mastery*, *16*(03). Retrieved from https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

Chan, J. C. W., & Paelinckx, D. (2008). Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, *112*(6), 2999-3011.

Chauhan, H., & Chauhan, A. (2014). Evaluating Performance of Decision Tree Algorithms 1. *International Journal of Scientific and Research Publication*, 4(4), 1-2.

Defiyanti, S., & Pardede, D. L. (2010). Perbandingan kinerja Algoritma ID3 dan C4. 5 dalam klasifikasi spam-mail. *Skripsi Program Studi Sistem Komputer*. Depok: Universitas Gunadarma.

Duda, R. O., Hart, P. E., & Stork, D. G. (1995). *Pattern Classification and Scene Analysis* (2nd ed.). New York: John Wiley & Sons, Inc.

Google Developers. Classification: Precision and Recall. *Machine Learning Crash Course*. Retrieved October 3, 2018, from https://developers. google.com/machine-learning/crash-course/ classification/precision-and-recall.

Google Developers. Classification: ROC and AUC. *Machine Learning Crash Course*. Retrieved October 3, 2018, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques* (Vol. 12). Springer Science & Business Media.

Hastuti, K. (2012). Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif. *Semantik*, *2*(1), 241-249. Retrieved from http://publikasi.dinus.ac.id/index.php/semantik/article/view/132

Iskandar, D., & Suprapto, Y. K. (2013). Perbandingan akurasi klasifikasi tingkat kemiskinan antara algoritma C4. 5 dan Naïve Bayes Clasifier. *JAVA Journal of Electrical and Electronics Engineering*, *11*(1).

ames, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: with Applications in R (Springer Text in Statistics)*. Springer.

Janardhanan, P., & Sabika, F. (2015). Effectiveness of Support Vector Machines in Medical Data Mining. *Journal of Communications Software and Systems*, 11(1), 25-30.

Karyadiputra, E. (2016). Analisis Algoritma Naive Bayes Untuk Klasifikasi Status Kesejahteraan Rumah Tangga Keluarga Binaan Sosial. *Technologia: Jurnal Ilmiah*, *7*(4), 199-208.

Kataria, A., & Singh, M. D. (2013). A review of data classification using k-nearest neighbor algorithm. *International Journal of Emerging Technology and Advanced Engineering*, *3*(6), 354-360.

Kaur, G., & Oberai, E. N. (2014). A review article on Naive Bayes classifier with various smoothing techniques. *International Journal of Computer Science and Mobile Computing*, *3*(10), 864-868.

Kini, M.M., Devi, S.H., G Desai, P., Chiplunkar, N. (2015). Text Mining Approach to Classify Technical Research Documents using Naive Bayes. *International Journal of Advanced Research in Computer and Communication Engineering*, 4(7).

Li, M., Ma, L., Blaschke, T., Cheng, L., & Tiede, D. (2016). A systematic comparison of different object-based classification techniques using high spatial resolution imagery in agricultural environments. *International Journal of Applied Earth Observation and Geoinformation*, *49*, 87-98.

Luque, I. F., Aguilar, F. J., Álvarez, M. F., & Aguilar, M. Á. (2013). Non-parametric object-based approaches to carry out ISA classification from archival aerial orthoimages. *IEEE Journal of selected topics in applied earth observations and remote sensing*, *6*(4), 2058-2071.

Mitchell, R. S., Michalski, J. G., & Carbonell, T. M. (2013). *An Artificial Intelligence Approach*. Berlin: Springer

Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*, *18*, 60.

Noviyanto, H. (2015). Pengklasifikasian Laman Web Berdasarkan Genre Menggunakan URL Feature. *Seminar nasional Teknologi Informasi dan Komunikasi*.

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). Decision trees: an overview and their use in medicine. *Journal of medical systems*, *26*(5), 445-463.

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, *97*, 205-227.

Quinlan, J.R. (1986). *Induction of Decision Trees. Machine Learning*, 81-106. Kluwer Academic Publishers.

TNP2K. Tentang Basis Data Terpadu. Retrieved September 24, 2018 from http://bdt.tnp2k.go.id/tentang/.

Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion Matrix-based Feature Selection. *MAICS*, *710*, 120-127.

Waikato University. Weka 3: Data Mining Software in Java. Retrieved September 27, 2018, from https://www.cs.waikato.ac.nz/ml/weka/.

YouTube. (2017). Random Forest - Fun and Easy Machine Learning. Retrieved September 25, 2018, from https://www.youtube.com/watch?v=D_2LkhMJcfY.

YouTube. (2017). Support Vector Machine (SVM) - Fun and Easy Machine Learning. Retrieved September 25, 2018, from https://www.youtube.com/watch?v=Y6RRHw9uN9o.

YouTube. (2013). Weka Tutorial 28: ROC Curves and AUC (Model Evaluation). Retrieved October 3, 2018, from https://www.youtube.com/watch?v=j97h_-b0gvw.