Identify Level of Welfare Population Based on Income Levels Using Decision Tree Method

Yunita Ardilla[*1], **Wilda Imama Sabilla**[2], **Sarah Astiti**[3]
[*1]Department of Da'wah Management, Universitas Islam Negeri Sunan Ampel Surabaya
[2]Department of Information Technology, Politeknik Negeri Malang
[3]Department of Information System, Institut Teknologi Telkom Purwokerto
e-mail: *[1]yunita.ardilla@uinsby.ac.id, [2]wildaimama@polinema.ac.id, [3]sarah@ittelkom-pwt.ac.id

Abstract

Identification of population welfare influenced by several factors. This identification is useful to assist the government in classifying the level of welfare population which is useful for providing subsidies to be targeted. Therefore this study aims to determine the level of welfare population based on the level of income per capita using decision tree method. The selection of the best model is based on the calculation value of accuracy, precision, and recall with k-fold cross validation method. Based on experiments that have been done, it can be concluded that the decision tree model produced has good performance with a tree shape model has 622 leaves with tree size 705 of nodes, the model has an accuracy of 86,97%, precision 0.897 and recall 0.917.

Keywords—*Classification, Decision Tree, Prosperty Level*

1. INTRODUCTION

Indonesia's population in 2021 will reach 270 million people [1]. The large number of residents raises various increasing demands for needs. The welfare that is aspired to has not been felt evenly. It seems the number of poor people in 2021 reached 27.54 million [2]. The government has a responsibility to improve the welfare of the community. The community will be more prosperous if the government services are getting better and more evenly distributed. These services include providing various goods or services to be provided to the community.

The process of giving government subsidies, it is often found that there are subsidies are not on target. Sometimes the subsidies that should be received by the underprivileged instead are received by those who are able. This very detrimental to the government. So the government usually does a review to identify the recipient of the subsidy based on the level of income per capita. Identification of per capita income level is influenced by several factors, including based on age, hourly wages, length of time someone has worked, education level and so on [3]. Through these factors the government can find out the level of welfare population per capita. The identification of the level of welfare population per capita can be classified by the decision tree method.

Decision tree has been advantages of being simpler and specific, easier to interpret, and more flexible in choosing features from internal nodes. However decision tree has several drawbacks like including overlapping between classes and accumulating errors. In terms of performance, the decision tree is more good compared to other classification algorithm [4].

The purpose of making this paper is to create a model that is able to classify the level of welfare population based on the level income per capita. The model created by decision tree method with software WEKA.

2. METHOD

The data used is the Census-Income (KDD) dataset from UCI Datasets, with the following information:

Table 1. Census-Income (KDD) Datasets Information from UCI Datasets

Amount of data	**199523**
Number of attributes	42 (including class attribute)
Attribute information	Age Class of worker Detailed industry recode Detailed occupation recode Education Wage per hour Enrolled in Edu inst last wk Marital status Major industry code Major occupation code Race Hispanic origin Sex Member of a labor union Reason for unemployment Full or part time employment stat Capital gains Capital losses Dividends from stocks Tax filer status Region of previous residence State of previous residence Detailed household and family stat Detailed household summary in household Instance weight Migration code-change in msa Migration code-change in reg Migration code-move within reg Live in this house 1 year ago Migration prev res in sunbelt Num persons worked for employer Family members under 18 Country of birth father Country of birth mother Country of birth self Citizenship Own business or self employed Fill inc questionnaire for veteran's admin Veterans benefits Weeks worked in year Year Class

Dataset contains of population census data, then the datasets is grouped into two categories, namely residents with incomes below US$50.000 (-50.000) and residents with incomes above US$50.000 (50.000+). In this dataset the sum of each data for the population class is as follows:

Table 2. Total Population Class Data on Dataset

Class of population	Amount
-50.000	187141
50.000+	12382

In the case of identification level welfare population used the decision tree method. Decision tree is a simple classification method. The algorithm used is C4.5. In making a decision tree, several stages need to be done. The first stage is preprocessing which includes cleaning up data from missing values, then selecting attributes and discretizing continuous data. Attribute selection useful for removes irrelevant and redundant attributes. Discretization also need to useful for change numeric data into nominal / categorical data because decision tree can only be used for nominal data [4]. The next stage is to create a model tree from training data. Furthermore, the model that has been made from training data will be measured by performing tests using data testing.

2.1 Selection Attributes

Selection attribute is used to select attributes from the dataset to reduce the dimensions of the data. Selection attribute is done by removing redundant attributes and attributes that are not relevant or contain information that is not needed for data mining purposes. The selection attribute algorithm used is the best-first search algorithm with the following steps:

1. The set N becomes a sequential list of initial nodes (initial nodes)
2. If N is empty, exit and give the message failure.
3. The set N becomes the first node in N, and removes n from N
4. If N is the destination node/goal then exit and give a success message.
5. In addition, add n to N, rank the nodes in N according to the estimated distance of goal, and return to step 2.

2.2 Discretization

Discretization is used to convert numerical data into nominal data in the process of building a model tree. In this case, the discretization algorithm used is Fayyad & Iran’s MDL Method [5], the steps are:

1. Sorts values on attributes
2. Look for potential cut-points. Cut points are point on the sorted attributes which the class label changes (for example class changes is from class A and Class B
3. Evaluate information gain based on a particular method (information gain, gain ratio, gini coefficient, chi-squared test) at each cut-point and select the largest value.
4. Repeat until the values change no more.

2.3 Algoritma C4.5

This algorithm developes ID3 method to build decision tree model from training data. Training data is set *S = s1, s2, ..., sn* samples that have been classified. Every samples *si = x1, x2, ..., xn* is vector where *x1, x2, ..., xn* is an attribute of feature of the sample. This training data is added with vectors *C = c1, c2, ..., cn* which is *c1, c2, ..., cn* is the class of each sample. Tree model creation is based in entropy information from several possible tree shapes. The attribute with the highest information gain is chosen to be added to tree model. The steps continue until all the attributes are arranged in the tree model. Pseudocode from C4.5 algorithm is as follows [6]:

1. Check *base case*
2. For each attribute *a*, look for information gain that has been normalized from the splitting of attribute *a*.
3. *a_best* is an attribute with the highest information gain value.
4. Make *decision node* the separates *(split) a_best*
5. Repeat the *sublist* obtained from the *a_best* then add the node as *child* of the node.

2.4 Calculate Information Gain

Information gain is used to select attributes that will be used as nodes in the process of building the model tree. To calculate information gain, the entropy calculation must be done first. With the following functions:

$$Entropy(t) = -\sum_{j} p(j \mid t) \log p(j \mid t) \qquad (1)$$

While the function to calculate information gain is as follows:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right) \qquad (2)$$

2.5 Prunning

Pruning is advantage of the C4.5 algorithm compared to othes tree compilation methods. With pruning the risk of errors in classification process can be reduces. Namely by specialized training data. Thereby making trees more general.

2.6 Testing The Classifier Model

Classifier model trials are conducted to determined the level of accuracy of the model in terms of classifying the data inputted. The trial is conducted with the following scenario:

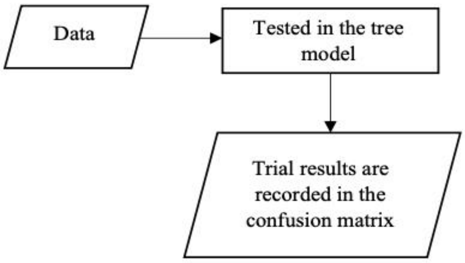

Figure 1. Flow Chart Tree Model Trial Scenarios.

The k-fols cross validation method is used in this testing process. This method divides the dataset into k subsampling, k-1 subsamples are used as training data and a sample is used as testing data. The results of the trial are recorded in a confusion matrix containing the results of predictions made by the model. After the confusion matris is formed, performance calculations can be done in three ways, namely the calculation of accuracy, precision, and recall.

$$\text{Precision (p)} = \frac{a}{a + c}$$

$$\text{Recall (r)} = \frac{a}{a + b} \qquad (3)$$

$$\text{Accuracy} = (a + d / a + b + c + d)$$

Description:
a = true positive
d = true negative
b = false negative
c = false positive.

3. EXPERIMENTAL RESULT

Testing algorithm in this study using WEKA software, WEKA provides 4 kinds of testing methods [7]. Before testing the algorithm, the first steo is to preprocessing data by removing missing values and removing redundant data. Then make the attribute selection, attribute selection using the best-first search algorithm.

The dataset attributes which initially amounted to 42 attributes, the reduced to 11 attributes including: Education, Major Occupation Code, Sex, Capital Gains, Capital Losses, Dividen from Stocks, Tax Filer Stat, Instance Weight, Family Member Under 18, Weeks Worked in Year, and Class. Because the classification process will be carried out then the data that has been selected will be discretized, the discretized process changes a number of numerical data into categorical data. The following are the attributes that have gone through the process of discretization is Capital Gains, Capital Losses, Dividens from Stocks, Instance Weight, Weeks Worked in Year.

The results of the decision tree method for classifying the level welfare population obtained a tree that has 622 leaves, with a tree size of 705 nodes. The resulting tree model is in the form of a multiway split. Next, the testing process is carried out on the model tree that has been generated to measure the performance of the model. In this paper used k-fold cross validation for the test method. Here are the results of testing tree model that is formed:

Table 3. Performance Measurement Result Table Using K-Fold Cross Validation Method.

k-fold cross validation		**Performa Model**		
	Class	**Accuracy**	**Precision**	**Recall**
10	50000+	86,905%	0,803	0,765
	-50000		0,897	0,916
11	50000+	86,893%	0.802	0.766
	-50000		0.897	0.915
12	50000+	86,873%	0.803	0.763
	-50000		0.896	0.916
13	50000+	86,96%	0.805	0.764
	-50000		0.897	0.917
14	50000+	86,898%	0.803	0.765
	-50000		0.897	0.916
15	50000+	86,973%	0.805	0.764
	-50000		0.897	0.917
16	50000+	86,970%	0.805	0.764
	-50000		0.897	0.917

Description:
50000+: class of prosperous population level

-50000: class of less prosperous population level
After testing with the k-fold cross validation method, look for the average of each iteration of *k*. The following results from the calculation of the average of each iteration.

Table 4. Table of Average Calculation Results for Each Iteration.

K-Fold Cross Validation	Accuracy	Precision	Recall
10	86,905%	0.868	0.869
11	86,893%	0.868	0.869
12	86,873%	0.867	0.869
13	86,96%	0.868	0.869
14	86,898%	0.868	0.643
15	86,973%	0.868	0.87
16	86,970%	0.68	0.87

From this table it can be seen that *k* = 16 the tree model has the best performance because it has better accuracy and recall than others.

4. CONCLUSION

Based on experiments conducted, namely building a decision tree model for classifiying the level welfare population, then obtained a tree that has 622 leaves, with a tree size of 705 nodes. The resulting tree is in the form of multiway split. The tree model has the best performance when testing with k-fold cross validation with k = 16.

5. SUGGESTIONS

The suggestions for further research are:

1. Expected to try compare other classification methods besides the decision tree.
2. The application of data mining for population census data processing with the decision tree method is a process to generate new knowledge in the form of comparisons between the factors that affect the population census data.
3. The results of data mining using the decision tree method is an arrangement of sequences of activities that support each other in the process of classifying the level of population welfare so that it is easier to understand by looking at the stages of the decision tree image.

THANK-YOU NOTE

The author would like to thank the editors of the JUPITER journal who have supported the author's opportunity so that this research article can be published.

BIBLIOGRAPHY

[1] Hasil Sensus Penduduk 2020," Badan Pusat Statistik, 21 January 2021. [Online]. Available: https://www.bps.go.id/pressrelease/download.html?nrbvfeve=MTg1NA%3D%3D&sdfs=ldjfdifsdjkfahi&twoadfnoarfeauf=MjAyMS0wOC0xNiAxNDozNToyMQ%3D%3D. [Accessed August 2021]

[2] Persentase Penduduk Miskin Maret 2021 turun menjadi 10,14 persen," Badan Pusat Statistik, 15 July 2021. [Online]. Available: https://www.bps.go.id/pressrelease/2021/07/15/1843/persentase-penduduk-miskin-maret-2021-turun-menjadi-10-14-persen.html. [Accessed August 2021]

[3] Terran lane and Ronny Kohavi. Census-Income (KDD) Data Set. U.S. Census Bureau. http://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29

[2] Kumar, V, "Introduction to Data Mining", in Person Addison Weasly Publishing Company, 2006.

[3] Gorunescu, Florin, "Data Mining: Concepts and Techniques", Verlag Berlin Heidelberg: Springer, 2011

[4] Wei, W, "Time Analysis Univariate and Multivariate Methods," in Addison Wesley Publishing Company, 2006.

[5] Sathyadevan, S., and Remya, R, "Comparative Analysis of Decision Tree Algorithms: ID3, C4.5 and Random Forest," in Computational Intelligence in Data Mining, pp 549-562, 2014.

[6] Kirkby, R., Frank, E., and Reutemann, P, "WEKA Explorer User Guide for Version 3-5-7", 2007

[7] Kusumadewi, S and Guswaludin, I., 2005, Fuzzy Multi Criteria Decision Making, Media Informatics, Vol. 3 No.1., 25-38. [8] Kurniawan, E., Mustafidah H., and Shofiyani, A., 2015, TOPSIS Method to Determine New Student Admissions for Medical Education at Muhammadiyah University of Purwokerto, JUITA ISSN: 2086-9398, Vol. 3, No.4.