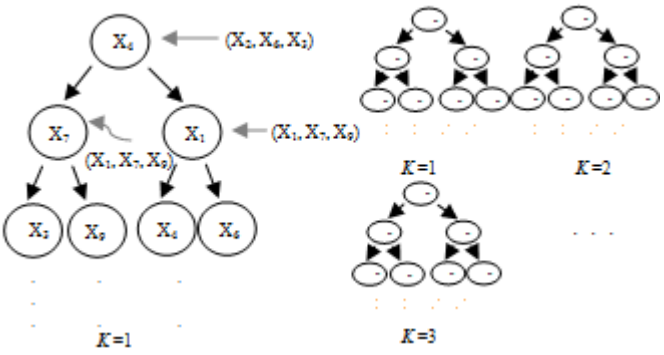
get the final classification prediction [6]. RF is developed in 2001 by Leo Breiman from the process of bagging. If in bagging process is used resampling bootstrap to generate the classification tree in many version, then, combine to get the final prediction, so in RF, the random process tp form the classification tree is not only done by data sample, but also by collecting the predictor variable. So, this process will produce a set of classification tree with different size and form [7].



Source : Bae & Bickel, 2009

Figure 2: (a) the illustration of single tree construction (b) the illustration of Random Forests [8]

The RF construction is in figure 2. While, RF algorithm is started by taking *n* data sample from the first dataset by using resampling bootstrap technique by reversion, then, arranging the classification tree from every dataset of resampling bootstrap result without the pruning process, by determining the best splitter based on the random predictor variables. Then, doing classification prediction of sample data based on the formed classification tree. Reply the stages so it is gained some desired classification tree. The reply is done in *k* times. Then, the next, it is done the classification prediction of final sample data by combining the prediction results of classification tree based on the *majority vote* rules.

2.3 The Measurement of Classification Accuracy

The accuracy of classification results can be measured by Apparent Error Rate (APER) and total accuracy rate (1-APER). In Table 1 shows the cross tabulation to calculate the classification accuracy.

Table 1. The Classification Table

Observation	Prediction		Total
	1	2	
1	*n11*	*n12*	*n1.*
2	*n21*	*n22*	*n2.*
Total	*n1.*	*n2.*	*N*

$$\text{Apparent Error Rate} = \frac{\text{the number of wrong prediction}}{\text{the number of total prediction}} = \frac{n_{21} + n_{12}}{N}$$

$$\text{Total accuracy rate} = 1 - APER$$

Table 2. The Research Variables

Variable	Variable Name	Category
Y	The expected assistance package for household	1: Housing 2: Modal 3: Cattle 4: Cash Money 5: Daily Needs 6: Education / Health
X_1	The large of floor building of the house	-
X_2	The type of floor building of the house (the largest)	1: Ceramic/marbles/granite 2: floor tile/tile/terrazzo 3: Cement/red brick 4: Wood/Board 5: bamboo 6: Land
X_3	The type of the wall building of house (the largest)	1: wall 2: wood 3: Bamboo
X_4	The Status of Ownership of defecation the bowel facility (privy/WC)	1: personal ownership 2: collective ownership 3: Public Ownership 4: there is no
X_5	The source of drink water	1: water in packages 2: plumber water 3: Pomp water 4: well 5: water source 6:river water
X_6	The place for disposing of fecal water	1: Septic tank 2: pond/field 3: river /damn 4: land hole 5: open area /garden
X_7	The source of main lighting	1: PLN electrical with metrical 2: PLN electrical without metrical (joining/ others) 3: electrical not from PLN 4: not electrical
X_8	The fuel for cooking	1: electrical 2: Gas/LPG 3: kerosene 4: wood charcoal /coconut shell 5: wood
X_9	The frequency of daily eating	-
X_{10}	The frequency of meat/milk/chicken meat consumption in a week	-
X_{11}	The frequency of buying new cloth in a year	-
X_{12}	The place for healing	1: hospital/the center of public health /sub-health center 2: doctor 3: paramedics 4: traditional
X_{13}	The final education certificate of the head of family	1: no certificate 2: elementary school 3: junior high school 4: high school 5: Diploma I/II 6: higher than academy
X_{14}	The number of family assets owned	-
X_{15}	The type of roof of the house (the largest)	1: concrete 2: roof tile 3: wood 4: iron sheeting 5: Asbestos 6: palm fiber/sago palm
X_{16}	Monthly income	-
X_{17}	The status of house ownership	1: personal ownership 2: contract 3: rental 4: free-rent 5: official house 6: parents or relatives
X_{18}	The large of house building	-

2.4 The Poverty Concept

BPS and Social Department stated that the poor household of the households under the poverty line is the households with individual incapability of economy aspect to meet the minimum basic need for the proper life [9]. Based on BPS, the government implement some stages in solving the proverty problems [10]. These stages are realized in three packages of assistance programs, namely:

1. Package I is social assistance and protection. This assistance package is for protection and fulfilment the rights of education, health, food, sanitation, and clean water. This assistance program is realized in the form of rise for poor households (Raskin), Society Health Insurance (Jamkesmas), School Operational Assistance (BOS), Dreamed Family Program (PKH), and Cash Direct Assistance (BLT).
2. Package II is society emporwement (National Program of Society Empowerment-Independent). This package is to give protection and fulfilment the rights of participation, work and business chances, land, Natural resources and Life Environment, and also Housing.
3. Package III is the Empowerment of Micro and Small Business (UMK-KUR) which the goal is to protect and fulfil the rights of work and business chances, natural resources and life environment.

3. The Research Methodology

3.1 The Data Sources

The data is from secondary data gained from the results of Verification Survey of Poor Household in Jombang Regency in 2011. This survey is done by Regional Development Planning Agency (Bappeda) in Jombang regency with the goal to verify the result survey of BPS about the condition of poor households.

3.2 The Research Variables

The research variables used are response variable (Y) and predictor variable (X) showed in Table 2. The predictor variables are health, education, social, economy and human resources aspect. While, the response variable (Y) is the expected assistance package for poor households which is classified into six groups. They are housing assistance like house renovation, electrical installment, giving the cash assistance (BLT), the assistance for daily need (rise for poor households, free main daily needs, food assistance and others), giving the fund for child education (BOS) and Poor Student Assistance (BSM) and also the health insurance (Society Health Insurance and Insurance for Old People).

4. Result and Discussion

4.1 The Descriptive Statistics

The number of RTSM and RTM in Jombang Regency in 2011 is 11.763 and 21.108 households. The data from Bappeda used in this research after pre-processing data is 6.362 RTSM and 15.932 RTM. The distribution of poor households in Jombang regency based on the expected assistance package for household is 0,094 percent of the RTSM and 0,087 percent of the RTM willing the housing assistance (house renovation, electrical, and others), and 0,227 percent of RTSM and 0,198 RTM willing the cash modal and good modal, and 0,303 percent of RTSM and 0,285 percent of RTM requiring cattle, and 0,307 percent of RTSM and 0,350 percent of RTM requiring cash money assistance, and 0,040 percent of RTSM and 0,058 percent of RTM willing the daily needs like the nine-basic needs or food assistance, 0,029 percent of RTSM and 0,022 percent of RTM willing the education and health assistance.

4.2 The CART Analysis

The Formulation of Maximum Classification Tree

The tree formulation is started by splitting all possibilities of splitter variables and threshold by using the Gini Index. The splitter and threshold with the highest goodness of split value will be the best splitter. The tree formulation is done so it is formed he maximum tree and by the end there is no splitting anymore. So, it is formed the maximum tree with high number of terminal node.

The example of syntax R for tree formulation of RTSM classification is shown as follow,

```
>library (rpart) #load the library
>rtsm<-read.table("E:/rtsm.txt",header=TRUE) #load the data
>kontrol<-rpart.control(minbucket=10,cp=0,xval=10)
>fit<-
rpart(y~.,data=rtsm,parms=list(split="gini"),method="class",control
=kontrol) #Grow tree
>printcp(fit) #display the results
>summary(fit,file="D:/summaryrtsm.txt") #detailed summary of
splits
```

The result of CART analysis state that the maximum classification tree gained has 35 depth level with the number of terminal nodes are 271 nodes for RTSM classification. While, the maximum classification tree for the RTM classification result has 60 depth level with the number of terminal nodes are 692 nodes.

The Pruning Of Maximum Classification Tree

To avoid the under/over fitting case and to ease the analysis process of classification tree, it is done the pruning by using 10-fold cross-validation estimate method. The pruning is stated in “control” syntax, that is (xval=10) at time of formulation the classification tree.

The Pruned Of Optimal Classification Tree

The pruning result with 10-fold cross-validation estimate method is evaluated based on the error (cost) value gained. The pruning produced the minimum error is selected as the optimal tree. The Syntax used is as follow,

```
>fitp<-prune(fit,cp=fit$cptable[which.min(fit$cptable[,"xerror"]),
"CP"]) #prune the tree
>#plot the pruned tree
>plot(fitp,uniform=T,compress=T,branch=1,margin=0.01,main="Pr
uned Classification Tree for RTSM")
>text(fitp,use.n=T,all=TRUE,cex=.6)
```

The classification tree for the selected RTSM is the optimal tree with 22 nodes of terminal nodes, the complexity parameter value is 0,001588, the minimum error is 0,851827, and the relative error is 0,820967. While, the selected of RTM clssifictaion tree is 28 nodes of terminal nodes, the value of complexity parameter is 0,000917, the minimum error is 0,898166, and the relative error is 0,870753. The optimal classification tree for RTSM and RTM is shown in Figure 3.

There are 18 predictor variables involved in research being the tree classification former. And based on the construction of optimal classification tree, it is known that the variable as the parent node in the classification is the type of fuel for cooking for RTSM and the monthly income for RTM. Those variables are the main splitter and the most determination in classifying households based on the expected assistance packages.

Figure 3: (a) The Construction of Optimal Classification Tree for RTSM (b) The Construction of Optimal Classification Tree for RTM

The Result of CART Classification Tree Accuracy

The evaluation of CART classification properness is done by predicting the data research using the results of classification tree gained. The example of syntax R used to predict and calculate the classification properness is as follow,

```
>#prediction and calculate 1-APER of dataset
>fitp.pred1<-predict(fitp,rtsm,type="class")
>write.csv(fitp.pred1,"D:\\predictionrtsm.csv")
>crosstab1<-table(fitp.pred1,rtsm$y[-13830])
>APER1<- 1-(sum(diag(crosstab1))/sum(crosstab1))
>accuracy1<-(sum(diag(crosstab1))/sum(crosstab1))
```

The accuracy level of the optimal classification tree result for RTSM and RTM can be seen in Table 3. The result in Table 3 is the RTSM classification accuracy showed by the total accuracy rate (1-APER) gained by 0,4313 with 0,5687 classification error level (APER). While, the RTM classification accuracy is total accuracy rate (1-APER) by 0,4338 with 0,5662 of classification error level (APER).

Table 3. The Optimal Classification Tree Accuracy

Classification	APER	1-APER
RTSM	0,5687	0,4313
RTM	0,5662	0,4338

4.3 The RF-CART Analysis

The classification with *random forests* (*RF*) uses random sample data to formulate the tree; and in every splitting, there is limitation to the number of random predictor variable [11]. Thus, in this research, the determined control parameter is random predictor variable to be used as the tree former in every splitting with tree variables. Besides, the number of formed classification tree (CART) is 300 trees. The example of syntax R for RTSM classification with RF-CART is as follow,

```
>library(randomForest) #load library
>fit.rf<-randomForest(y~.,data=rtsm,mtry=3,ntree=300,
importance=TRUE) #grow tree for RF
>importance(fit.rf) #importance of each predictor
>varImpPlot(fit.rf) #visualize importance results
>print(fit.rf) #view results
```

The RF-CART results states that the variables with higher influence in RTSM classification into six packages of household assistance respectively are monthly income, the place for disposing fecal water, the type of floor building, and the source of drinking water variables. While, the most important variables in RTM classification are monthly income, the place for disposing fecal water, the source of drinking water and the large of building house.

If there is prediction and calculation for the classification properness by using syntax R as follow,

```
>fitrf.pred1<-predict(fit.rf,rtsm[-13830,])
>write.csv(fitrf.pred1, "D:\\pkm\\rfrtsm.csv")
>crosstabrf1<-table(predicted=fitrf.pred1,observed=rtsm[-
13830,"y"])
>APERrf1<-1-(sum(diag(crosstabrf1))/sum(crosstabrf1))
>accuracyrf1<-(sum(diag(crosstabrf1))/sum(crosstabrf1))
```

Then, the accuracy result will be seen in Table 4.

Table 4: The RF-CART Classification Accuracy

Classification	APER	1-APER
RTSM	0,0050	0,9950
RTM	0,0167	0,9833

Based on the Table 4, it can be seen that RTSM classification accuracy with RF-CART shown by total accuracy rate (1-APER) is 0,9950 with 0,0050 of classification error level (APER). While, the RTM classification accuracy with RF-CART is 0,9833 of total accuracy rate (1-APER) with 0,0167 of classification error level (APER).

4.4 The Comparison of Classification Results Between CART and RF-CART

Table 5: The Comparison of Classification Result Between CART and RF-CART

Classification method		APER	1-APER	The increase of 1-APER
RTSM	CART	0,5687	0,4313	0,5637
	RF-CART	0,0050	0,9950	
RTM	CART	0,5662	0,4338	0,5495
	RF-CART	0,0167	0,9833	

The comparison of classification result between RTSM and RTM by using CART and RF-CART is shown in Table 5. The RF-CART research method has smaller APER value and bigger 1-APER value compared with CART method for both classification for RTSM and RTM. The increase of accuracy level by using ensemble random forest method is 0,5637 for RTSM and 0,5495 for RTM. Thus, it can be said that RF-CART method is the better method in the RTSM and RTM Classification in Jombang regency based on the expected packages of household assistance compared to the CART classification method.

4.5 The Prediction of RTSM and RTM Distribution In Jombang

If it is done the prediction of analysis data to all households in Jombang by using RF-CART classification results as the high accuracy in classification method, it can be gained the map about the most expected package of household assistance by poor households in every district in Jombang. The map is shown in Figure 4.

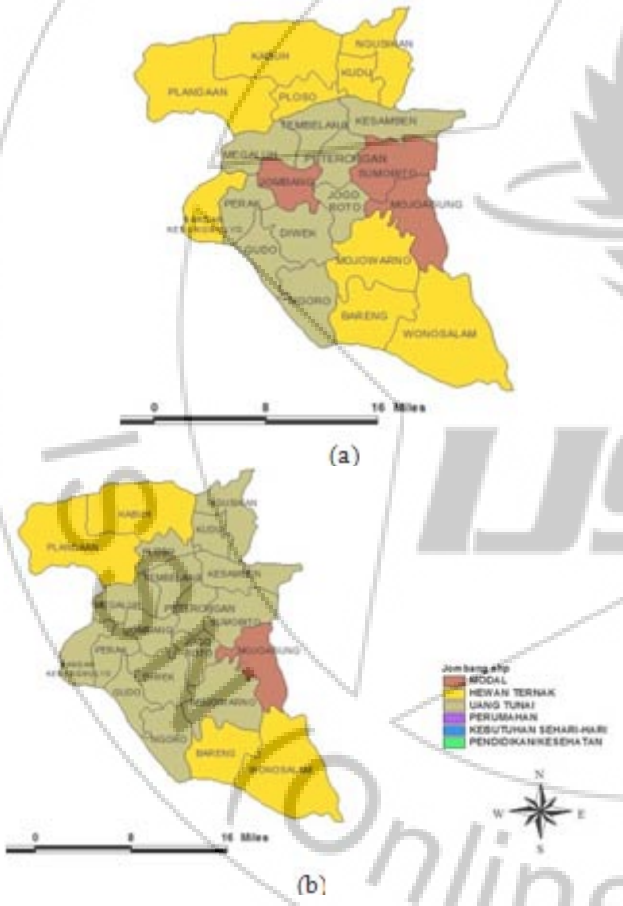


Figure 4: (a) The Map of Expected Assistance Packages by RTSM in Jombang (b) The Map of Expected Assistance Packages by RTM in Jombang

The majority of RTSM in Jombang, Sumobito, and Mojoagung want the assistance like modal, (loan/good modal), while the RTSM in nine district, which are Plandaan, Kabuh, Ploso, Kudu, Ngusikan, Bandar Kedungmulyo, Mojowarno, Bareng, and Wonosalam want cattle from government, and the rest expects the cash assistance like BLT and BLSM. While, the majority RTM in Mojoagung district expects modal assistance, RTM in Plandaan, Kabuh, Bareng, and Wonosalam district want cattle, while the rest expects cash money assistance.

5. Conclusion

Based on the analysis and discussion done, it can be concluded that:

1. RTSM and RTM classification based on the expected household packages by CART method produce 0,4314 dan 0,4338 of total accuracy rate. The classification result shows that the most important variable in tree formulation process and determination of the expected household packages is the fuel for cooking variable for RTSM classification and month income variable for RTM classification.
2. RF-CART classification can produce 0,9950 of total accuracy rate for RTSM classification and 0,9833 for RTM classification. The variables with important role in RTSM classification are monthly income, the place for disposing fecal water, the type of floor building and the source of drinking water. While, the most important variables in RTM classification are monthly income, the place for disposing fecal water, the source of drinking water and the large of building house.
3. The comparison result of classification accuracy between CART and RF-CART method is the RF-CART method gives the increase in 0,5637 of total accuracy rate for RTSM classification and 0,5495 for RTM classification, if it is compared with CART method. Thus, the RF-CART method is the better method in classifying poor households in Jombang based on the expected household packages.
4. The recommendation for government is to increase the provision of cattle, increase the foundation for the People Business Credit and distribute the cash money assistance for poor people who needs it.

References

[1] R. J. Lewis, “An Introduction to Classification and Regression Trees (CART) Analysis,” Annual Meeting of the Society for Academic Emergency Medicine, UCLA Medical Center, California, 2000.
[2] C. D. Sutton, “Classification and Regression Trees, Bagging, and Boosting,” Handbook of Statistics, 24, pp. 303-329, 2005.
[3] M. J. Muttaqin, “Metode Ensemble pada CART untuk Perbaikan Klasifikasi Kemiskinan di Kabupaten Jombang,” Thesis of Statistics Department, Institut Teknologi Sepuluh Nopember, Surabaya, 2013.
[4] A. A. Hidayanti, “Boosting Multivariate Adaptive Regression Spline (MARS) Binary Response untuk Klasifikasi Kemiskinan di Kabupaten Jombang,” Thesis of Statistics Department, Institut Teknologi Sepuluh Nopember, Surabaya, 2013.
[5] L. Breiman, J. Friedman, R. Olshen, & C. Stone, “Classification and Regression Trees,” Chapman Hall, New York, 1995.

[6] M. V. Wezel, & R. Potharst, R, "Improved Customer Choice Predictions using Ensemble Methods. European Journal of Operational Research", 181, pp. 436-452, 2007.
[7] A. Liaw, & M. Wiener, "Classification and Regression by Random Forests," R News, 2,pp. 18-22, 2000.
[8] C.Bae, & P. Bickel, "Versions of Random Forests: Properties and Performances," University of California, Berkeley (2009, Mar 29)
[9] BPS & Depsos, "Penduduk Fakir Miskin Indonesia 2002," BPS, Jakarta, 2002.
[10] BPS, "Analisis dan Penghitungan Tingkat Kemiskinan 2008," BPS, Jakarta, 2008.
[11] S. S. Shih, "Class 13a: Random Forests, for Model (and) Predictor Selection," Variation in Phonology, 251, p. 4, 2013.

Author Profile

Bambang Widjanarko Otok received the Dr., M.Si., and S.Si. degrees in Statistics from Institut Teknologi Sepuluh Nopember in 1994, Statistics from Institut Pertanian Bogor in 2000, and Matematics from Universitas Gadjah Mada in 2008, respectively. He now with Statistics Institut Teknologi Sepuluh Nopember, Surabaya.

Dian Seftiana received the S.Si. degree in Statistics from Institut Teknologi Sepuluh Nopember Surabaya in 2014. She has any interest about poverty, social, and nonparametric method.