Int. J. Advance Soft Compu. Appl, Vol. 10, No. 3, November 2018
ISSN 2074-8523

Identifying Determinant Factors to Internet Access Using Decision Tree

Wahyu Wibowo[1], I Nyoman Budiantara[2], and Bekti Cahyo Hidayanto[3]

[1]Department of Business Statistics
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
e-mail: wahyu_w@statistika.its.ac.id

[2]Department of Statistics
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
e-mail: i_nyoman_b@statistika.its.ac.id

[3]Department of Information System
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
e-mail : bekticahyo@is.its.ac.id

Abstract

Internet access is an important aspect in the digital literacy. There is many advantages to be able to access the internet in term of education, economic and information. Unfortunately, there is a lack of internet access of many Indonesian people. Thus, this paper aims to identify determinant factors to internet access of Indonesian people. The used method is decision tree as classification method in the supervised learning and the data is from National Socio-Economic Survey, East Java Province. To develop the decision tree model, the target variable is having internet access or not, meanwhile the input variables are socioeconomic factors, i.e. mobile phone usage, computer usage, age, gender, education level, residential location, working status. Based on the best decision tree model, the selected variables that influence people to access the internet are, age, education level, computer usage and mobile phone usage. In term of these variables, the characteristic of people who will access the internet the first is who use the computer and the second is from senior or higher education, using mobile phone, and more than 36 years old. The accuracy of this decision rules is 88% for both training and testing datasets.

Keywords: *classification, decision tree, socio-economic factors, internet access*

1 Introduction

Information and communication technologies (ICT) has become a important and prominent sector in recent decades. It is not only connecting the people, but also driving economic growth and industrial revolution. Research by OECD (Organisation For Economic Co-Operation And Development) shows that ICT investment has contributed to growth and labour productivity in all OECD countries, especially in the United Stated [1]. This fact will encourage any country to develop their economic and industry through ICT investment. However, there is a gap in term of ICT performance among country in the world.

The growth of internet users in Indonesia is also increasing every year. These developments can be seen from the proportion of Internet users. In June 2017, Internet World Stats released the proportion of Indonesian Internet users 50.4%, which means 132 million of the total 263 million Indonesians are internet users. Indonesia is also included as the country with the highest penetration rate of the 5th rank after China, India, USA, Brazil. The high proportion of internet users in Indonesia is in line with the infrastructure development carried out by the government, but it has not been able to realize the dream of the country after the construction of infrastructure supporting internet running. Due to the fact that infrastructure development by the government in the western part of Indonesia until mid-2017 touched its progress rate of 74% did not affect the population in East Java Province in using the internet. Noted that the percentage of internet users in East Java Province in 2017 only 31.17% only.

As additional, based on *ICT Development Index* 2017 released by International Telecommunication Union, the United Nations specialized agency for information and communication technologies, Indonesia is ranked 111 from 176 country. This index consists of 11 indicators built from 3 sub-indices which include the progress and development of ICT Access infrastructure, ICT Use and ICT Skill. In term of ICT Use sub-indice, one of indicators is percentage of individuals using the internet. By this indicator, percentage of Indonesian people using the internet is 25.37%. Indeed, this percentage increases from year to year, in average increase 1.5% annually from 2000-2016. However, this is still below compared to other country. So, it is important to more boost the Indonesian people to use the internet. This imply that is urgent to identify what of determinant factors people to access the internet. Thus, for this purpose, this paper aim to study determinant factors people to internet access of Indonesian people by taking case study from East Java people. This study will be observational based on secondary data from National Socio-Economic Survey. The used factor is limited on socioeconomic charateristic of respondent that hypotetically will relate the internet acess. As additional, this study will be data mining in order to gain informatin that can be utilized to increase percentage people to access the internet.

2 Related Work

Previously some authors have studied factors related to internet access. The study in Saudi Arabia shows that the internet usage is influenced most strongly by the number of educated people, the number of mobile subscribers, income, the number of fixed lines, and employment level [2]. This study utilize macro approach and apply regression model using time series data from 1994 to 2014. Another study is in rural population of Cyprus using logit model, shows that the age, the educational level, the income, the type of agricultural, activities and the location of the farm, are significant determinants for the personal computer and the Internet usage [3].

Another study was conducted to analyze the process of internet diffusion across the world using a panel of 214 countries during the period 1990–2004. The result is that the degree of competition in the provision of the internet contributes positively to its diffusion [4]. Meanwhile, study in China analyze relationships between factors in the framework EPIC (Economy-Policy-Infrastructure-Content) and internet diffusion using multivariate time series analysis. This study shows that the growth of Internet penetration was found to be mainly driven by Internet content and access cost; however, GDP per capita and telecommunications infrastructure were inconsequential [5].

The previously publication on information technology in Indonesia also had been done. Multiresponse semiparametric regression for modelling the effect of regional socio-economic variables on the use of information technology in East Nusa Tenggara [6]. More specific, the response variables are percentage of households has access to internet and percentage of households has personal computer. Then, predictor variables are percentage of literacy people, percentage of electrification and percentage of economic growth. Based on identification of the relationship between response and predictor variable, economic growth is treated as nonparametric predictor and the others are parametric predictors. Another research was held in rural area, villagers of Melung Village, Banyumas, Central Java. The technology adoption model test results using SEM PLS (Partial Least Square) shows that the demographic factors, social influence and facilitating conditions affect the acceptance and utilization of technology in rural areas [7]. Another research in villager Pasar VI, Kualanamu, Deli Serdang, North Sumatra show that factor of effort expectancy and social influences have significantly affected the intention to use the internet. This research use the model of the Unified Theory of Acceptance and Use of Technology (UTAUT) such as performance expectancy, effort expectancy, social influences and facilitating condition were analyzed as factors that can affect intention to use the internet [8].

3 Problem Formulations or Methodology

Although percentage of people who access internet increase from year to year, it is still below compared to the other country. There are many challenges how to increase the number of people to access the internet, such as social, demographic, culture and economic. Thus, it is not easy to identify which factor relate with people in accessing the internet as well as this will be aim of this work.

The data to support this work is from 2017 National Survey of Socio-Economic data. This survey is annually held by Statistic Indonesia across Indonesia region. However, in this study the data is only from East Java, one of provinces in Indonesia. This province was selected because it socioeconomic characteristic is representativeness of Indonesia. In term of the number of data, this study will involve a quite large of data, 96529 observation. This is different with the two previously works in Indonesia that is cited before, the data is limited from one village around hundreds.

More detailed, there are some variables from the survey hypothetically influenced the people to access the internet. Some of them are presented in Table 1. These variables are selected by considering the data availability and two cited previously about internet access in Indonesia.

Table 1: List of Variables

Variables (code)	Value	Type	Status
Internet Access	Yes, No	Categorical	target
Mobile phone usage	Yes, No	Categorical	Input
Computer usage	Yes, No	Categorical	Input
Age		Continuous	Input
Gender	Male, Female	Categorical	Input
Education level	No Education Primary School Junior High School Senior High School Higher Education	Categorical	Input
Residential location	Urban, Rural	Categorical	Input
Working Status	Work, Not-work	Categorical	Input

Next, by considering target and input variables as presented in Table 1, finding functional form between target and input is machine learning. There are so many methods in machine learning both unsupervised and supervised learning [9]. More specific, the problem of this study is a supervised learning or classification because the target variable has been defined. Classification is a vibrant research topic over the last decade era. There is 179 classifiers arising from 17 families (discriminant analysis, Bayesian, neural networks, support vector machines, decision trees, rule-based classifiers, boosting, bagging, stacking, random forests

and other ensembles, generalized linear models, nearest neighbors, partial least squares and principal component regression, logistic and multinomial regression, multiple adaptive regression splines and other methods) [10].

4 The Proposed Method

For this work, the algorithm will be used is decision tree, and is also called as Classification and Regression Tree (CART). This method was one of the classification methods or supervised learning [11,12]. Advantages of tree classification among others were:

1. This method was non-parametric so that it did not require binding assumption such as normal distribution assumption for predictor variable
2. Data structure could be visually seen so that it would be easy to explore and decide based on the model obtained.
3. It did not only provide classification but also give estimation of classification error probability
4. It enabled to identify interaction among predictor variables that were locally influential due to the application of gradual decision-making in a set of part of complex measurement data.
5. The final classification result was simply performed i.e. efficiently new data classification.
6. It made it easy to interpret the result

Currently, CART method is one of the prominent methods in data mining and machine learning especially in big data problem. CART method implementation was carried out with some stages i.e. determining training and testing data and construction of classification tree, pruning of a classification tree, determination of optimum classification tree.

- **Construction of classification tree**. Construction of tree was started with splitting all probabilities of splitter variables. Then with selection criteria of goodness of split which was calculated with Gini Index, selected splitter was splitter and threshold with the highest goodness of split. The tree constructed was called maximum tree which could not be further split afterward.
- **Pruning of classification tree**. To simplify the next analysis process, a maximum classification tree generated was pruning the tree using test sample estimate method. Each pruning result had a certain value of relative cost so that minimum relative cost value was selected.
- **Determination of Optimum Classification Tree**. An optimum classification tree is obtained from stage two. It is simpler and produce the simple rule. Class labelling had been carried out on each terminal node constructed in optimum classification tree.

The decision tree used a selective algorithm of binary recursive portioning. In the illustration of classification tree in Fig. 1, the very essential variable was called parent node containing all data with notation t_1. In Fig. 1 internal node is symbolized with hexagon i.e. t_2, t_3, t_7 whereas terminal node with rectangular t_4, t_5, t_6, t_8, and t_9. To calculate depth starts from first node t_1 at depth 1, whereas t_2 and t_3 at depth 2, and so on until terminal node t_8 and t_9 which are at depth 4.

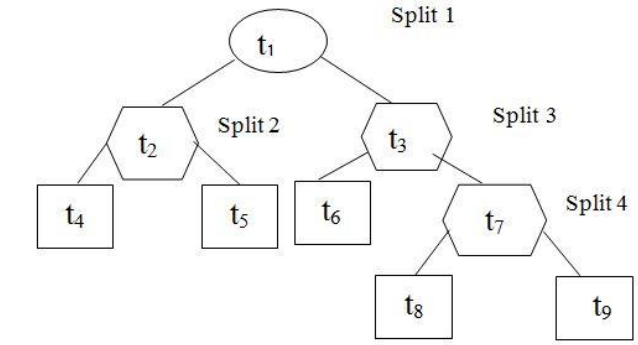

Fig.1 Decision tree illustration

The decision rules performance will evaluated by using accuracy and Receiving Operating Characteristic (ROC) Curve. If target variable is binary category, namely 0 and 1, accuracy is computed from confusion matrix as Table 1.

Table 1: Confusion Matrix

Actual Group		**Predicted Group**	
		1	**0**
Y	**1**	True Positive (n_{11})	False Positive (n_{10})
	0	False Negative (n_{01})	True Negative (n_{00})

$$Accuracy = \frac{n_{00} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}} \tag{8}$$

The ROC curve is curve created by plotting true positive rate and false positive rate and area under this curve is performance measurement of binary classifier. All of the computing process was implemented in Rattle (R Analytic Tool To Learn Easily), interface to conduct data science based on R environment [13].

5 Results, Analysis and Discussions

Graphical summary of the variables is presented in Fig. 2. In term of ICT usage, only 30% people have internet access, 18.8% have computer and 70.2% have mobile phone. Then, residence location of respondent 47% in the rural area and 53% in the urban area. Histogram of age shows that majority of respondents is

below 63 years old, about 90%. Percentage of male and female is 51% and 49% respectively. Meanwhile, the most education is from primary school, 42%, and then from senior and junior high school, 22% and 19% respectively. Then, the working status shows that 54% people is work and 46% is not-working.

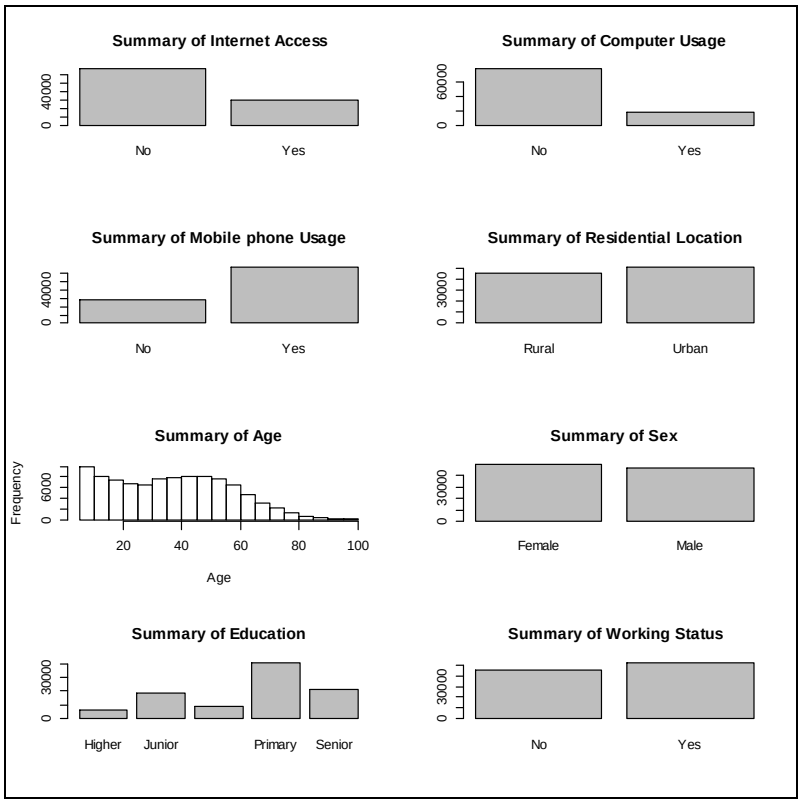

Fig. 2 Graphically summary of variables

Next, decision tree would be performed by Rattle. The first stage is to split the data become two groups, training set and testing set. The training set will be used to fit the model and will be validated using the testing set. The training set is selected 80% from the data and 20% as testing set. The second stage is to get the best decision tree model. The target and input variables in creating the decision model refer to Table 1. The optimum decision tree model is presented in Fig. 3. The meaning of each node and the detailed rule is in appendix. The result shows that the selected input variables involved in the decision tree are age, computer usage, education level, mobile phone usage.

From Fig. 3, the first node is the root node of the decision tree, numbered as node number 1 and refer to the target variable, internet acces. The information provided that the majority category for the root node is No, have 67570 observations and 69% correctly classified as No as well as 31% incorrectly classified as Yes. If the root node itself were treated as a model, it would decide

that it will be no internet access and based on the training dataset, the model would be 69% correct.

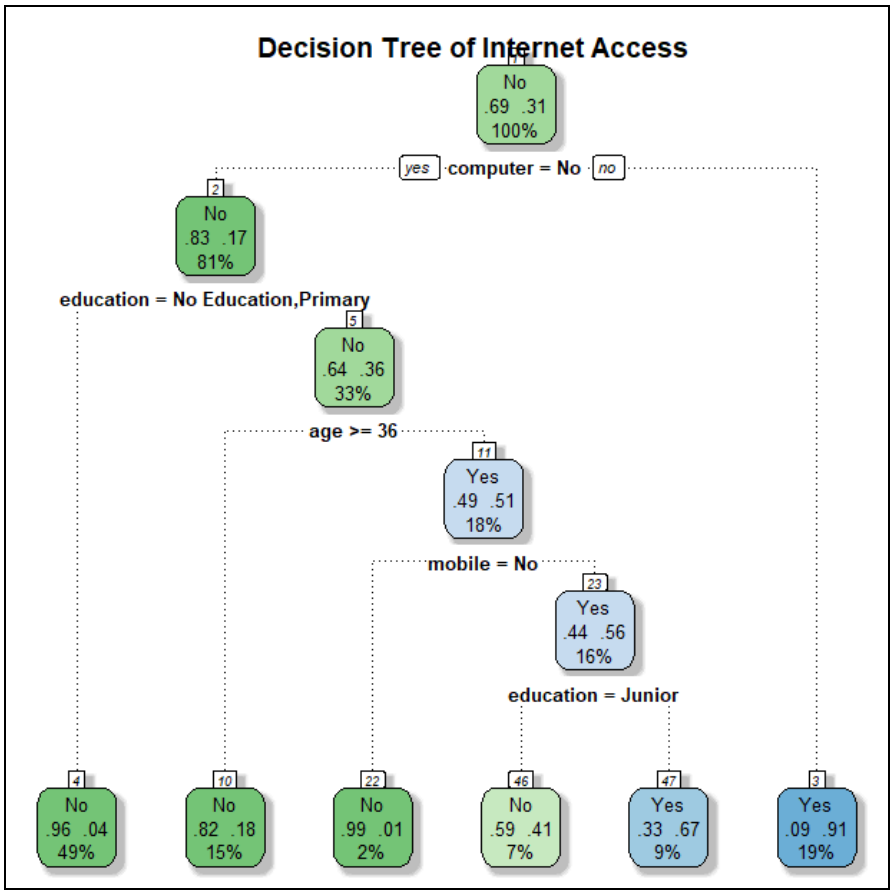

Fig. 3 Decision tree of internet access

Then, the root node is split into two sub nodes based on the variables computer, 81% observation for node number 2 and 19% for node number 3. Node 2 express No for computer usage and there is 54893 observations, 83% classified as No internet access. Node 3 express Yes for computer usage and there is 12677 observations, 9% classified as No internet access, 91% classified as Yes for internet access. Based on node 3, it can be inferred that the most people who has computer usage will access the internet with probability 0.91. Finally, the last terminal node is node 3.

After that, node 2 is split based on variables education. No-education and primary school category was combined to be node 4, as well as junior, senior, and higher education to be node 5. Note that node 4 is terminal node and 96% is classified as No internet access. This means that the tree is not split any further at node 4. However, node 5 is still split based on variable age, which cut off is 36 years old.

To be node 10 and node 11. Node 10 is terminal node for more than 36 years old and 82% classified as No internet access.

Node 11 is split based on variable mobile phone usage, to be node 22 and node 23. Node 22 is terminal node for No mobile phone usage which is 99% classified as No internet access. Node 23 is split based on variable education, to be node 46 and node 47. Node 46 is for junior education and it is terminal node. In addition, node 47 is also terminal node for both senior and higher education. For junior education, 59% classified as No internet access. On the contrary, for senior and higher education, 67% is classified as Yes internet access.

The two out of six terminal nodes are classified as Yes for internet access. Hence, the characteristic of people who will access the internet can be identified. The first is who use the computer and the second is from senior or higher education, using mobile phone, and more than 36 years old.

The third stage is to evaluate the decision tree model in term of predicting people whether will access internet or not based on the selected input variables for both training and testing set. For this purpose, the prediction of the target variable will be calculated by using the selected input variable and the rules. Then, by creating cross tabulation between predicted group and actual group of target variable, confusion matrix will be produced as presented in Table 2. Based on this table, accuracy is 88.04% and 87.86% for both training and testing data. These are a good level of accuracy.

Table 2: Confusion Matrix of Decision Tree Model

Actual Group		**Training Prediction**		**Testing Prediction**	
		No	**Yes**	**No**	**Yes**
Y	**No**	43994	2859	9456	632
	Yes	5218	15499	1126	3266

$$Accuracy\ (training\ data) = \frac{43994 + 15499}{43994 + 2859 + 5218 + 15499}\ x\,100\% = 88.04\%$$

$$Accuracy\ (testing\ data) = \frac{9456 + 3266}{9456 + 632 + 1126 + 3266}\ x\,100\% = 87.86\%$$

Furthermore, evaluation of the decision tree using area under ROC curve is presented in Fig. 4. Clearly that area under curve is 0.92 for both training and testing data. Evaluation of the optimum decision tree model show that the decision tree can show well that what factors influence the internet access. The factors are age, computer usage, education level, mobile phone usage.

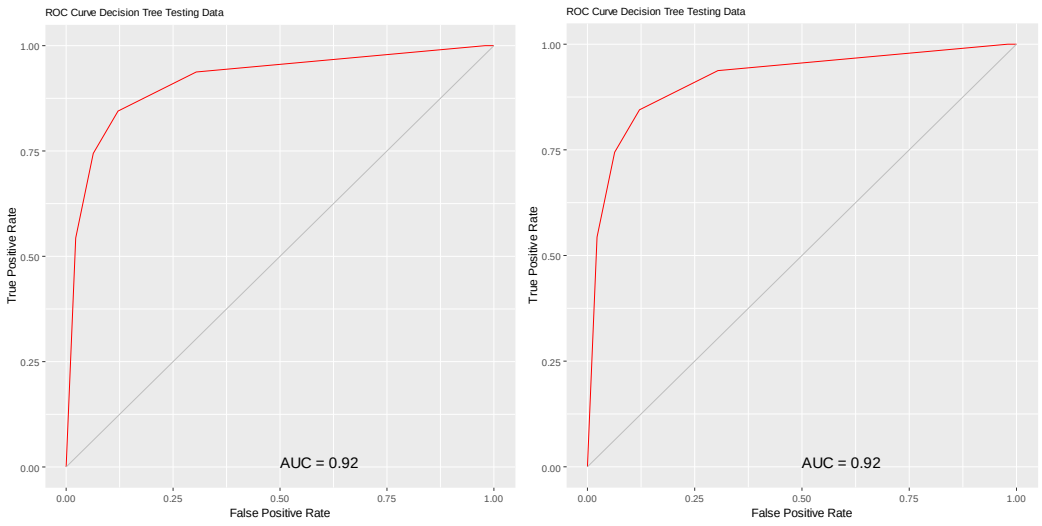

Fig. 4 ROC Curve

6 Conclusion

The decision tree method work well to identify determinant factor to internet access. This is indicated by a high accuracy and under curve of ROC curve, 88% and 0.92 respectively. The selected variables in the decision tree model are, age, education level, computer usage and mobile phone usage. In term of these variables, the characteristic of people who will access the internet the first is who use the computer and the second is from senior or higher education, using mobile phone, and more than 36 years old.

ACKNOWLEDGEMENTS

The authors are grateful to Institut Teknologi Sepuluh Nopember that supported this work in partly through Research Grant with contract number 1192/PKS/ITS/2018 (1302/PKS/ITS/2018. The authors also would like to thanks to Ryan Nurrahman for his effort to prepare the data from Statistics Indonesia.

References

[1]. OECD (2003), "The Contribution of ICT to Growth", in *ICT and Economic Growth: Evidence from OECD countries, industries and firms*, OECD Publishing, (pp. 35-54), Paris, http://dx.doi.org/10.1787/9789264101296-4-en.

[2]. Bardesi, H.J. (2016). Factors Affecting Demand For Internet Access In Saudi Arabia. *Eurasian Journal of Business and Management*. 4(3). 29-38. DOI: 10.15604/ejbm.2016.04.03.003

[3] Adamides, G., Stylianou, A., Kosmas, P. C., Apostolopoulos, C., D. (2013). Factors Affecting PC and Internet Usage by the Rural Population of Cyprus. *Agricultural Economics Review,* 14 (1), 16-36

[4] Andrés, L., Cuberes, D., Diouf, M., Serebriskya, T. (2010). The diffusion of the Internet: A cross-country analysis. *Telecommunications Policy*. 34 (5–6). 323-340, https://doi.org/10.1016/j.telpol.2010.01.003

[5] Feng, G.C. (2015). Factors affecting Internet diffusion in China: A multivariate time series analysis. Telematics and Informatics. 32(4). 681-693, https://doi.org/10.1016/j.tele.2015.02.009

[6] Wibowo, W., Wene, C., Budiantara, I.N., and Permatasari, E.O., (2017), Multiresponse Semiparametric Regression For Modelling The Effect of Regional Socio-Economic Variables On The Use of Information Technology, *AIP Conference Proceedings*, 1825, 020025; doi: 10.1063/1.4978994

[7]. Tambotoh, J.J.C., Manuputty, A.G., Banunaek, F.E., (2015),Socio-economics Factors and Information Technology Adoption in Rural Area, The Third Information Systems International Conference, *Procedia Computer Science*, 72, 178 – 185

[8] Susanto, A., (2015), "Factors Affecting The Behaviour Of Internet Use Of Villager Pasar VI, Kualanamu, Deli Serdang, North Sumatra", *Jurnal Penelitian Pos dan Informatika*, Vol.5 No 1 September 2015 : 65 – 86

[9] Hastie, T., Tibshirani, R., and Friedman, J., (2001), *The Elements of Statistical Learning*, Springer Series in Statistics Springer New York Inc., New York, NY, USA,

[10]Delgado, M.F., Cernadas, E., Barro, S., Amorim, D., (2014), Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?, *Journal of Machine Learning Research*, 15, 3133-3181,

[11]Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1993, Classification And Regression Tree. New York, NY: Chapman And Hall

[12] Roger J. Lewis. 2000. An Introduction to Classification and Regression Trees (CART) Analysis. Annual Meeting of the Society for Academic Emergency Medicine. California, UCLA Medical Center

[13]Williams, G. J. (2011), Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Use R!, Springer

Appendix : Decision Tree Rules

```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 67570

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 67570 20717 No (0.69339944 0.30660056)
   2) computer=No 54893  9093 No (0.83435046 0.16564954)
     4) education=No-Education,Primary 32995  1296 No (0.96072132 0.03927868) *
     5) education=Higher,Junior,Senior 21898  7797 No (0.64394009 0.35605991)
      10) age>=35.5 10547  1987 No (0.81160520 0.18839480) *
      11) age< 35.5 11351  5541 Yes (0.48815082 0.51184918)
        22) mobile=No 1023     7 No (0.99315738 0.00684262) *
        23) mobile=Yes 10328  4525 Yes (0.43812936 0.56187064)
          46) education=Junior 4647  1928 No (0.58510867 0.41489133) *
          47) education=Higher,Senior 5681  1806 Yes (0.31790178 0.68209822) *
   3) computer=Yes 12677  1053 Yes (0.08306382 0.91693618) *

Tree as rules:

 Rule number: 3 [internet=Yes cover=12594 (19%) prob=0.91]
   computer=Yes

 Rule number: 47 [internet=Yes cover=6048 (9%) prob=0.67]
   computer=No
   education=Higher,Junior,Senior
   age< 36.5
   mobile=Yes
   education=Higher,Senior

 Rule number: 46 [internet=No cover=4877 (7%) prob=0.41]
   computer=No
   education=Higher,Junior,Senior
   age< 36.5
   mobile=Yes
   education=Junior

 Rule number: 10 [internet=No cover=9997 (15%) prob=0.18]
   computer=No
   education=Higher,Junior,Senior
   age>=36.5

 Rule number: 4 [internet=No cover=33014 (49%) prob=0.04]
   computer=No
   education=No-Education,Primary

 Rule number: 22 [internet=No cover=1040 (2%) prob=0.01]
   computer=No
   education=Higher,Junior,Senior
   age< 36.5
   mobile=No
```