The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Using CART Model to Classify Family's Monthly Expenditure According to Urban and Rural Area in Sulaimani Governorate

Huda Mohammad Saeed[1], Aras Jalal Mhamad[2,3], Renas Abubaker Ahmed[4]

[1,2,4]Statistic & Informatics Dep., College of Administration & Economics, Sulaimani University, Sulaimani City, Kurdistan Region – Iraq
aras.mhamd@univsul.edu.iq[2]
[3]Accounting Dep., College of Administration & Economics, Human Development University, Sulaimani City, Kurdistan Region – Iraq

Abstract

The monthly family's expenditure is one of the important economic situation that appeared in the society especially in Sulaimani governorate in Kurdistan region – Iraq. The objective of this study is to classify family's monthly expenditures according to families who live inside (urban) and outside (rural) of Sulaimani city for year 2011 using the modern style in analyzing of classification which is (Classification and regression tree – CART) method, regression tree models are trained in a two-stage procedure, i.e. using recursive binary partitioning to make a tree structure, by a process of pruning to removing non-significant leaves, with the possibility of assigning multivariate functions to terminal leaves to improve generalization. CART was used to identify rank outcome explanatory s by determining monthly family's expenditures according to families who live in urban or in rural. The study found that the important variable is (Housing Rental) between eight variables for both rural and urban area at Sulaimani governorate, also of the 3210 families: 66.7% (2270) lived in Sulaimani city, and 33.3% (1131) lived in rural area in Sulaimani governorate according to housing rental variable, while 64.4% (1963) from families who lived in Sulaimani their housing rental is zero, and 87.0% (307) of families who lived urban area in Sulaimani city their housing renal between (201000 – 400000). Although the most families expenditure who live in Sulaimani city to food was ranged (151000 – 600000), household (101000 – 225000), education serves (101000 – 225000), health field (101000 - 150000), and transportation serves (251000 - 300000), while the most families expenditure who live in rural area in Sulaimani city for transportation serves (101000 – 150000), clothing (151000 – 200000), education serves (151000 – 225000), health field (76000 – 100000), and food (600000 – 750000).

Keywords: classification, CART model, statistical modeling.

الملخص

نفقات الأسرة الشهرية هي واحدة من الحالات الاقتصادية المهمة التي ظهرت في المجتمع وخاصة في محافظة السليمانية في إقليم كردستان - العراق. الهدف من هذه الدراسة هو تصنيف المستوى النفقات الشهرية للأسرة وذالك وفقًا للعائلات التي تعيش داخل المدن (المركز) وخارج المدن (الريف) في مدينة السليمانية لعام 2011 باستخدام الأسلوب الحديث (التصنيف و الانحدار الشجري - كارت)، عملية تجريب النماذج الانحدار الشجري يتم بمرحلتين ، أي يتم استخدام تجزئة ذو الحدين لإنشاء هيكلية الشجرة ، من خلال عملية إزالة المستويات غير معنوية ، مع إمكانية بناء دالة متعدد المتغيرات للوصل إلى حالة التعميم. باستخدام CART تم التعرف على رتبة توضيحية للنتائج من خلال تحديد نفقات

DOI: http://dx.doi.org/10.25098/3.1.9

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

الأسرة الشهرية وفقًا للعائلات التي تعيش في داخل المدن أو الريف. وجدت الدراسة أن أهم متغير هو (تأجير المساكن) بين ثمانية متغيرات لكل من المناطق الريفية و داخل المدن في محافظة السليمانية ، أيضًا من 3210 أسرة: 66.7٪ (2270) يعيشون في مدينة السليماني ، و 33.3 ٪ (1131) يعيشون في المناطق الريفية في محافظة السليمانية وفقا لمتغير تأجير المساكن ، في حين أن 64.4 ٪ (1963) من العائلات التي كانت تعيش في السليمانية إيجارها السكني هو صفر ، و 87.0 ٪ (307) من الأسر التي تعيش في داخل المدن في مدينة السليمانية سكنهم الكلي ما بين (201000 - 400000) دينار. على الرغم من أن معظم نفقات الأسر التي تعيش في مدينة السليمانية على الغذاء تراوحت ما بين (151000 - 600000) دينار، الأسرة (101000 - 225000) دينار، يخدم التعليم (101000 - 225000) دينار، المجال الصحي (101000 - 150000) دينار، ويخدم النقل (251000 - 300000) دينار، في حين أن معظم إنفاق العائلات الذين يعيشون في المناطق الريفية في مدينة السليماني للنقل يخدم (101000 - 150000) دينار، والملابس (151000 - 200000) دينار، والتعليم (151000 - 225000) دينار، والمجال الصحي (76000 - 100000) دينار، والغذاء (600000 - 750000) دينار.

مفاتيح الكلمات- التصنيف، النماذج CART، النمذجة الاحصائية

پوختە

خەرجی مانگانەی خێزان بەیەکێك لە بواری گرنگەکانی ئابووری دادەنرێت لەکۆمەڵگادا بەتایبەت لە پارێزگای سلێمانی هەرێمی کوردستانی عێراق. ئامانج لەم تۆێژینەوەیە بریتیە لە پۆلێنکردنی خەرجی مانگانەی خێزان، بۆ ئەو خێزانانەی لەناوەوەو دەرەوەی شاری سلێمانیدا ژیاون لە ساڵی 2011 بەبەکارهێنانی شێوازێکی تازەی لێکۆڵینەوە کە بریتیە لە ڕێگای پۆلێن و چەماوەی لاری (CART). مۆدێلی چەماوەی لاری بە دوو قۆناغ جێبەجێ دەکرێت، بەکارهێنانی دابەشکردنی دووانی بۆ دروستکردنی شێوەی درەختێك، بە پرۆسەی هەرەسهێنانی ئەو گەڵایانەی کە گرنگ نین لادەبرێن، لەگەڵ تەرخانکردنی مۆدێلی فرە گۆڕاو بۆ گەڵاکانی کۆتایی بەمەبەستی بەرەوپێشبردنی گشتگیرکردن. CART بەکارهێنراوە بۆ پلەبەندی شیکردنەوەی ئەنجامەکان بەدیاریکردنی خەرجی مانگانەی خێزان بۆ ئەو خێزانانەی کە لە ناوەوەو دەرەوەی شاری سلێمانی دەژین. تۆێژینەوەکە دەریخست کە گرنگترین گۆڕاو بریتیە لە کرێ خانوو لە نێوان هەشت گۆڕاودا بۆ هەردوو ناوچەکانی ناوەوەو دەرەوەی پارێزگای سلێمانی، هەروەها بۆ 3210 خێزان: لە سەدا 66,6 (2270) یان لەناو شاری سلێمانی و لەسەدا 33,3 (1131) یان لەدەرەوەی شاری سلێمانیدا ژیاون بە پێی گۆڕاوی کرێی خانوو،لەکاتێکدا لەسەدا 64,4 (1963) ی ئەو خێزانانەی لەناو شاری سلێمانیدا ژیاون خاوەنی خانووی خۆیانن وە لەسەدا 87 (307) ی ئەو خێزانانەی لەدەرەوەی شاری سلێمانی ژیاون کرێ خانوویان لەنێوان (201000-400000) دینار هەزاردا بووە. لەگەڵ ئەمانەشدا ئەو خێزانانەی لەناو شاری سلێمانیدا ژیاون زۆربەی خەرجی مانگانەیان بۆ خۆراك بووە کە لە نێوان (151000-600000) دینار، وە بۆ پێداویستی ناوماڵ لەنێوان (101000-225000) دینار، بۆ خوێندن لەنێوان (101000-225000) دینار، بۆ کەرتی تەندروستی لەنێوان (101000-150000) دینار وە بۆ خزمەتگوزاری گواستنەوە لەنێوان (251000-300000) دینار دا بووە. لەکاتێکدا زۆربەی خەرجی مانگانەی ئەو خێزانانەی لە دەرەوەی شاری سلێمانی ژیاون بۆ خزمەتگوزاری گواستنەوە لەنێوان (101000-150000) دینار، بۆ جلوبەرگ لەنێوان (151000-200000) دینار، بۆ خوێندن لە نێوان (151000-225000) دینار، بۆ کەرتی تەندروستی لەنێوان (76000-100000) دینار وە بۆ خواردن لەنێوان (600000-750000) دینار دا بووە.

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

1.1 Introduction

Economic Expenditures have a big effect on family economics, and these effects make changes of society economics, for example, permanent diseases effect on family expenditure such as arthritis of rheumatoid, which in return will be a heavy burden on family's economical state [20]. In addition, restaurants increase a part of family expenditures [8]. Many times increasing expenditure results in poverty, which endangers family health and wellbeing; these conditions could result in a gradual deterioration of poor family's health, which is to increase the burden on systems of health in the future [26]. While limited expenditure decreases a part of long drawn expenditures, such as newer of attached buildings to be decreasing household energy expenditure [24]. For these reasons classifying and organizing monthly family expenditure requires performing a scientific and academic method, one verified model is classification regression tree (CART) that is used for this purpose. CART model are widely used, e.g., in biomedical, educational, behavioral, psychological, and social sciences [3, 31, 16, 19]. CART model is a recursive method of partitioning, which builds classification and regression trees for predicting continuous or categorical explanatory variables (classification) and either continuous nor category dependent variables (regression) [10]. The CART is made through splitting subsets of the dataset by using all explanatory variables to construct two child nodes repeatedly, and the finally to produce homogeneous subsets of the dataset with respect to the target variable [18]. So that the aim behind this work is to classify family monthly expenditure according to their location of living in and outside Sulaimani city; via using CART model.

1.2 Research Question

To classify family's monthly expenditure between urban and rural Area in Sulaimani Governorate, this study has formulated the following set of questions:

1. What is the extent of the family's monthly expenditure between Urban and Rural Area?
2. What are the major factors that affect classify the family's monthly expenditure between Urban and Rural Area?
3. How can classify the family's monthly expenditure between Urban and Rural Area?

1.3 Objective of the Study

The main objectives of this study are:

1. To classify family's monthly expenditures of Urban and Rural Area in Sulaimani Governorate.
2. To examine the major factors that affect classify the family's monthly expenditure between Urban and Rural Area.

DOI: http://dx.doi.org/10.25098/3.1.9

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

1.4 Hypothesis of the Study

The following hypotheses have been empirically tested to answer the research questions:

$\mathbf{H_0}$: Family expenditures of Urban and Rural Area in Sulaimani city have not a similar classification.

$\mathbf{H_1}$: Family expenditures of Urban and Rural Area in Sulaimani city have a similar classification.

1.5 Literature Review

Bae H., Olson B. H., Hsu K. & Sorooshian S**.,** they predict for bacterial concentrations by using CART model for beach closure management. The results of their study showed that the all bacteria explained a different tree while there are some significant variables for each of them, the dissolved oxygen variable had an important effect for both total and fecal coliform. RMSE between 5 and 6.5% [2]. Abdul Kareem S., et al. used CART model to predict a continuous response variables and categorical or explanatory variables. They found that CART were predict the survival of AIDS by accuracy model between 60-93% depend on selected response variables, and they proved their result with a high Receiver Operating Characteristics (ROC) [1]. Iliev I. P., et al, used the CART method to construct a tree with binary regression. The resulting CART tree takes into account which input quantities influence the formation of classification groups and in what manner [14]. Pouliakis A., et al, they evaluated CART model for triage rules production and estimate of cervical intraepithelial neoplastic (CIN) risk in cases with ASCUS+ in cytology. The CART build a tree with inadequate cytological cases outcome and there are a high diagnostic accuracy of ancillary techniques. The CART performance was better than any other single test involved in this study [22]. Zimmerman R. K., et al, used CART model to estimate probabilities of influenza. They concluded that the CART of their study had a high sensitivity and NPV, while it had low PPV for detecting influenza [34].

Also, Cheng Z. MD MPH, et al, explored whether a knowledge–discovery approach building a CART prediction model for weight loss (WL) in the patient treatment of head and neck cancer (HNC) with radiation therapy (RT). Among 391 patients identified, weight loss explanatory s during radiation therapy planning were international classification of diseases diagnosis; with some other factors, and low-dose planning target volume–larynx distance were significant predictive factors. The ROC curve from their results during RT and EOT was 0.773 and 0.821, respectively [5]. And Yoo K., et al, applied a CART model with multiple linear regression (MLR) for predicting of potential urban airborne bacterial hazards according to AD events by using met genomic analysis and real-time qPCR. They showed that the CART method had more successfully predicted potential airborne

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

bacterial hazards with a high determination coefficient (R^2) and small bias, with the lower RMSE and MAE when comparing with the MLR method [32].

2. Materials and Methods

2.1 Classification and regression tree (CART)

The CART model is a method of non-parametric statistical approach that can merge both scale and categorical variables into analysis. Moreover. In a classification problem, we are given a set of data of training data. Each entry has a variables number. There is one distinguished variable called the response variable and the remaining variables are referred to as the independent variables. The main goal of using CART analysis is to construct the models which will use the independent variables for predicting the values of the dependent variable in the method of non-parametric [17]. Although, the CART analysis is a process of partitioning method of binary recursive data set depend on a criterion of specific splitting. The CART procedure is used an intensive algorithm which is searches to the best split between all possible split points for each independent variable in order to produce classification or regression model. To select the best split, can be use goodness-of-split criteria to evaluate the reduction in "impurity" was achieved by the optimal partitioning and then all splits are ordered based on their impurity reduction. However, where the field is missing for the best split, the next best split will be used, which is termed as the alternative split. The graphical output of the CART analysis resembles an inverted tree with internal and terminal nodes [7, 17].

In the extreme, one may grow a tree so large that each terminal node contains only one entry. Like a perfectly tree may classify the training data, while will most likely occur big errors in the testing data and predictions, which is referred to as "over fitting". CART makes a procedure of pruning-tree by means of either a testing data set after a large tree is grown or cross-validation to avoid over fitting. With the pruning, the analysis goes to an optimal tree which is used to prediction. In the optimal tree, each terminal node is correlated with a set of "rules" which deal with a sequence of splitting criteria that goes to the formation of that specific node. The rules are important because; first, they are used for predicting the dependent variable value. Second, they contain a wealth of information about the relationship between the dependent and the independent variables and the interactions among the independents [17].

2.2 Development and test of CART

The CART is one of the tree-based classification models that are the non-parametric models. Fig. 1 explains the development and test of CART procedure, which consists of data recursive splits into a set of subgroups representing categorical variables. The example of Fig. 1 explains the datasets can be split into two subgroups which are "Good" and "Poor" through internal nodes in diamond shape having splitting variables, *A*, *B*, *C*, *D*

DOI: http://dx.doi.org/10.25098/3.1.9

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

and split points, *a*, *b*, *c*, *d* . The terminal node in rectangular shape represents subgroups. In the CART, the splitting variables and the split points for each step are selected through the process of training, and using the Gini index to conduct this process as given in Eq. (1):

$$Gini = \sum_{k=1}^{K} \hat{P}_{mk}(1 - \hat{P}_{mk}) \quad \dots\dots\dots (1)$$

Where ˆ*mk p* is the proportion of subgroups *k* in the node *m*. The Gini index decreases when the particular subgroup proportion is large, the splitting variable and point showing Gini index with lowest value were chosen. The chosen split were continued until the tree was fully grown [27].

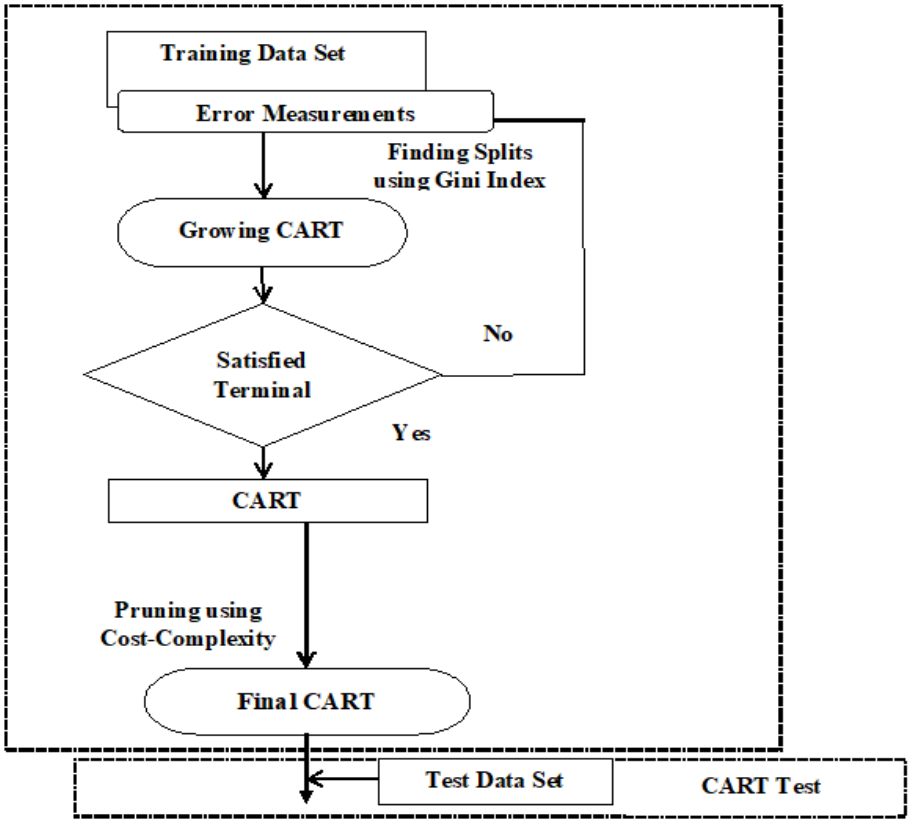

Figure1. CART Development and test Procedure

2.3 CART building

The process of tree building begins by splitting the root node into two child nodes. CART finds the best split by considering all probable splits for each response or explanatory variable. The best split is obtained when the function of impurity is minimized, which exists between the parent node and two child nodes. In CART, coefficient of Gini is used to measure the "purity", and can be defined as equation (2):

$$G(t) = 1 - \sum_{j=1}^{k} p^2\,(j|t) \dots\dots.. (2)$$

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Where t, k and p are the node, the categories number in the variables output, and the probability when the sample output variables take the j of probability for the node t. It reaches its minimum (zero) when all cases in the node fall into a single target category.
CART uses coefficient of Gini algorithm to reduction measure of the heterogeneity, and it's mathematically defined by:

$$\Delta G(t) = G(t) - \frac{N_r}{N} G(t_r) - \frac{N_l}{N} G(tl) \dots\dots.(3)$$

where $G(t)$ and N are the coefficient of Gini to the variables output and sample size before grouping, $G(tr)$, Nr, $G(tl)$ and Nl are respectively the coefficient of Gini and sample size of right sub tree, the Gini coefficient and the sample size of the left sub tree after grouping. Can be obtain the division point whose decreasing the heterogeneity fastest by repeating the method. In the regression tree (continuous response variable), strategy of selecting the grouping variable (optimal) is the same as the classification tree, and the main difference is variance as the measure indicator for output variable heterogeneity. Its mathematical definition is:

$$R(t) = \frac{1}{N-1}\sum_{i=1}^{N}(y_i(t) - \bar{y}(t))^2 \quad \dots\dots\dots\dots(4)$$

Where $t, N, y_i(t)$ and $\bar{y}(t)$ are the node, the sample size for the node t, the output in a variable's value for t, and the average of the output variables for the node t. Therefore, the measure of heterogeneity decreasing is variance reduction, its mathematical definition is:

$$\Delta R(t) = R(t) - \frac{N_r}{N} R(t_r) - \frac{N_l}{N} R(tl) \quad \dots\dots\dots\dots(5)$$

where $R(t)$ and N are the variance of output variable and sample size before grouping, $R(t_r)$, N_r, $R(tl)$ and N_l are respectively the variance and sample size of right subtree, the variances and the sample size of the left sub tree after grouping. To achieve maximum of $\Delta R(t)$ variable should be the best grouping variable. The method for determining the best point of division is the same as the classification tree [33].

2.4 CART Pruning

Due to the completion tree on the training samples feature is described too accurate, it loses general representation and cannot be used in the classification or prediction of new data, which is called over-fitting. To solve this problem can be use pruning technique, at the same time the accuracy of tree decision increases by 25%.
CART uses the method of minimum cost complexity pruning as pruning algorithm, that is an inspection and distinct process, also can be calculate the accuracy prediction of the current decision tree for the sample of test in the process of pruning; finally get an optimal tree with the balance of complexity and error rate. This method relies on a complexity

DOI: http://dx.doi.org/10.25098/3.1.9

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

parameter, denoted as α, which is gradually increased during the pruning process. The error of decision tree is referred as the cost, and the complexity of the decision tree T can be expressed as:

$$R_\alpha(T) = R(T) + \alpha \left|\tilde{T}\right| \quad \dots\dots\dots\dots (6)$$

where R(T), $\left|\tilde{T}\right|$and α are classification error on testing sample set, the number of leaf nodes, and the complexity for each increasing of the leaf node, respectively [33].

2.5 Determine Optimal Tree

The maximum tree will always fit the learning dataset with good accuracy than any other tree. The performance of the maximum tree on the original learning dataset, termed the "substitution cost again", generally greatly overestimates the performance of the tree on an independent set of data obtained from a similar population. This happen because the maximum tree fits properties and noise in the learning dataset, which are not probably occur with the same pattern in a different set of data. The goal in selecting the optimal tree, defined with respect to expected performance on an independent set of data, is to find the correct complexity parameter a so that the information in the learning dataset is fit but not over fit. In general, finding this value for require an independent set of data, but this requirement can be avoided using the technique of cross validation. The figure to the right shows the relationship between tree complexity, reflected by the number of terminal nodes, and the decision cost for an independent test dataset and the original learning dataset. The nodes increases numerally, the cost of decision decreases for the learning data. This corresponds to the fact that the maximum tree will always produce the best fit to the learning dataset. In contrast, the expected cost for an independent dataset goes to a minimum, and then increases as the complexity increases. This reflects the fact that an over fitted and overly complex tree will not perform well on a new set of data [23].

2.5.1 Cross Validation

Cross validation methods in CART are used to determine the optimal tree size for each run. The procedure of cross validation is based on optimal proportion between the complexity of the tree and misclassification cost. With the increase in size of the tree, misclassification error is decreasing and in case of maximum tree, misclassification error is equal to zero.

Letting $R(T)$ be the resubstitution estimate of the misclassification rate of a tree, T and $|T|$ be the number of terminal nodes of the tree, for each $\alpha \geq 0$ the cost–complexity measure, $R_\alpha(T)$, for a tree, T , is given by

$$R_\alpha(T) = R(T) + \alpha|T| \quad \dots\dots\dots\dots (7)$$

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Here, $|T|$ is a measure of tree complexity, and $R(T)$ is related to misclassification cost. α is the contribution to the measure for each terminal node. To minimize this measure, for small values of α, trees having a large number of nodes, and a low re-substitution estimate of misclassification rate, will be favored. For large enough values of α, a one node tree will minimize the measure. CART proceeds by dividing the learning sample into 10 nearly equal parts; each contains a similar distribution of the target variable. CART takes a part of data which are first nine parts, then builds the biggest possible tree, and uses the remaining of data to estimate the error rate of selected sub-trees initially. The same process is repeated on another 9/10 of the data while using a different 1/10 part as the test sample. The process continues until each part of the data has been held in reserve one time as a test sample. The results of the 10 mini-samples are combined to form error rates for trees of each possible size. The default method in CART is 10-fold cross validation. Usually an attempt is made to keep the same class frequencies in each fold, and each fold is stratified by the outcome variable of interest. This ensures that a similar distribution of outcome is present in each of the k subset of data [25]. Taking each of $\alpha_k(\alpha_m, \alpha_{m-1}, \dots, \alpha_2, \alpha_1)$ from the sequence of complexity parameters, obtain the optimal subtree, $T_{(-1),k}$ of $T_{(-1)}$ corresponding to αk. Then, one would have a sequence of the optimal sub tree of $T_{(-1)}$, that is $T_{(-1),m}, T_{(-1),m-1}, \dots, T_{(-1),2}, T_{(-1),1}$.

Thus, the final cross validation estimate, $R^{CV}(T_{\alpha k})$ of $R(T_{\alpha k})$ follows from averaging $R^{ts}(T_{(-1),k})$ over $i = 1, 2, \dots, 10$.

$$R^{CV}(T_{\alpha k}) = \frac{R^{ts}(T_{(-1),k}) + R^{ts}(T_{(-2),k}) + \cdots + R^{ts}(T_{(-10),k})}{10} \quad \dots\dots\dots(8)$$

2.5.2 Standard Error of R^{cv}

The sub-tree corresponding to the smallest $R^{cv}(T_{ak})$ is obviously desirable. There are some issues that arise when dealing with the cross-validation estimate (CV cost). The cross-validation estimates generally have substantial variability due to the random number generator used to separate the dataset into k test sets [25]. And oftentimes, when choosing the "right-sized" tree with the minimum CV cost, there will be several trees with CV costs close to the minimum. Breiman et.al make a reasonable suggestion that one should choose the "right-sized" tree with the smallest sized (least complex) tree whose CV cost does not differ appreciably from the minimum CV cost. Breiman proposed a revised strategy to select the final tree, which takes into account the standard errors of the cross validation estimates, the so called "1 SE rule". For making a selection of trees with CV costs close to the minimum, Breiman suggests to choose as the "right-sized" tree the smallest-sized tree whose CV cost does not exceed the minimum CV cost by more than one standard error of the CV cost for the minimum CV cost tree.

DOI: http://dx.doi.org/10.25098/3.1.9

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

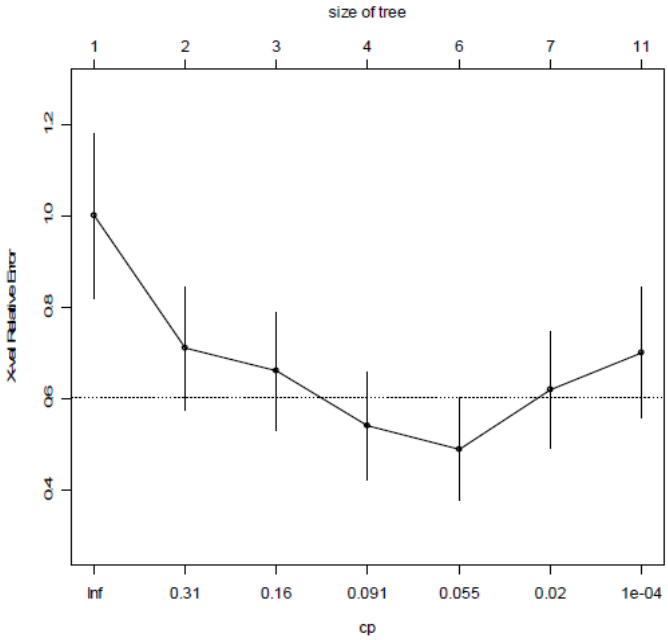

Figure2: Graphical representation of complexity parameter.

The graph above illustrates the relation between complexity parameter and the tree complexity (the size of tree refers to the number of the terminal nodes) to the cross validation estimate (x-val Relative Error). Vertical lines (error bar) represent the standard error around the Rcv for the different tree complexities. The horizontal line refers to the minimum of CV cost plus one standard error. As shown in the graph above, the CV costs, approach the minimum as tree size initially increases, and start to rise as tree size becomes very large. Note that the selected "right-sized" tree is close to the inflection point in the curve, that is, close to the point where the initial drop on CV costs with increased tree size starts to level out [25].

2.6. Model assessment

To assess performance of the models, can be use the receiver operating characteristic (ROC) curve tool. The ROC curve is built using sensitivity as the Y-axis correspond 1-specificity as the X-axis with various cut-off points as shown equation (9) [12]. The area under the ROC curve (AUC) explains the model capability to predict landslide and non-landslide pixels. When AUC value is equal to 1, then indicates a perfect model, while an AUC zero value indicates a worse model [28], and a higher AUC value indicates a better predictive capability of a model. According to [29], correlation of predictive capability and AUC could be quantified as follows: excellent between (0.9–1), very good (0.8–0.9), good (0.7–0.8), average (0.6–0.7), and poor (0.5–0.6) [4, 6, 11, 13]

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \ldots\ldots\ldots (9)$$

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Where true positive is denoted by TP and true negative is denoted by TN, TP and TN are the pixels number which are correctly classified, while false positive is denoted by FP and false negative is denoted by FN, they are the pixels numbers which are incorrectly classified [9, 30].

3. Data Analysis and Results

3.1 Data Description

Data were obtained from a Directorate of Statistics in Sulaymaniyah, Kurdistan-Iraq. 4013 observation were used in the study. The dataset used for the analysis is contained 23 variables and deals with monthly expenditure family in urban and rural at Sulaimani province for year 2011. The explanatory variables were utilized in this study as follows: costing of food (Food) which ranged between (less than 150000 through more than 1000000), transportation serves which ranged between (less than 50000 to more than 350000), costing of Clothing (Clothing Cost) which ranged between (less than 50000 to more than 350000), health serves which ranged between (less than 25000 to more than 225000), communication cost which ranged between (less than 25000 to more than 225000), education cost which ranged between (less than 25000 to more than 225000), household costing which ranged between (less than 25000 to more than 225000), and house rental which ranged between (less than 200000 to more than 400000). In the other hand the target variable is Living place (Yi), which is consist of two level; either Yi = 1 if Living place is classified as ‘‘urban’’, or Yi = 0 if Living place designation is ‘‘rural’’.

3.2 CART output

The CART method algorithm solves the classification and regression problem [17]. CART is a nonparametric solution tree technique which builds classification or regression trees depending on whether the indictor variables which are numerical. In our case. The algorithm is used to build a binary solutions tree. The goal is to find a tree which allows for a good distribution of the data with the lowest possible relative error of prediction [17]. More specifically, the aim of the method of regression tree is to data distribution in relatively homogeneous end nodes and to get a mean observed value at each node in the form of a predicted value as shown in Fig.3. When building regression trees, the validation is usually applied, since they may be sensitive with random errors in the data. This validation technique allows the construction of reliable models to standard regression models in CART. In this study we used 10% V-fold cross-validation, which is means the data are divided into 10 equal subgroups randomly, each of them containing 10% of the

DOI: http://dx.doi.org/10.25098/3.1.9

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

dataset. In order to select the tree and it's to reverse prune so as to find a tree with an optimal relative error for the data, we apply the standard cross-validation procedure as shown in Fig.3 and table 1:

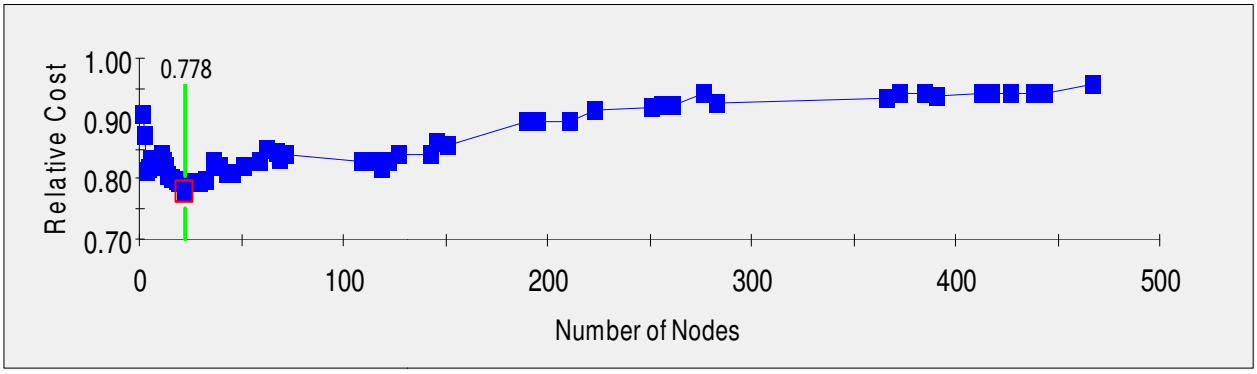

Figure3: Explain relationship between relative errors with number of nodes

Fig.3 explains the relationship between classification error and size of the tree, that is labelled a relative cost curve, and it is always ranged between zero and one, zero means no error or perfect fit model, while one represents the performance of random guessing, in this study the best tree is that with 22 nodes, and the relative error is (0.778), this tree decreases the percentage of incorrect classification in the training sample, and obtains suitable performance in the validation sample, the tree sequence is as shown in Table 1.

Table 1: Relative error for all possible tree (Tree Sequence)

Terminal Tree	Nodes	Test Set Relative Cost	Re-substitution Relative Cost	Complexity Parameter
1	467	0.95730 +/- 0.04283	0.33909	0
44**	22	0.77802 +/- 0.04194	0.69059	0.001602
45	20	0.79248 +/- 0.04182	0.69752	0.001741
46	19	0.79729 +/- 0.04173	0.70101	0.001758
47	16	0.79938 +/- 0.04092	0.71326	0.002051
48	14	0.80679 +/- 0.04113	0.72301	0.002447
49	13	0.81894 +/- 0.04092	0.72826	0.002636
50	12	0.82872 +/- 0.04092	0.73486	0.003309
51	11	0.83857 +/- 0.04113	0.74151	0.003337
52	8	0.82102 +/- 0.03957	0.76206	0.003435
53	6	0.83124 +/- 0.04113	0.77721	0.003797
54	5	0.81861 +/- 0.03973	0.79515	0.008983
55	4	0.81354 +/- 0.03892	0.81398	0.009422
56	3	0.87195 +/- 0.04199	0.83474	0.010392
57	2	0.90486 +/- 0.02020	0.90543	0.035353
58	1	1.00000 +/- 0.00000	1	0.047295

The general topological structure of the resulting classification and regression tree with 22 nodes is given in Fig. 4.

DOI: http://dx.doi.org/10.25098/3.1.9

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

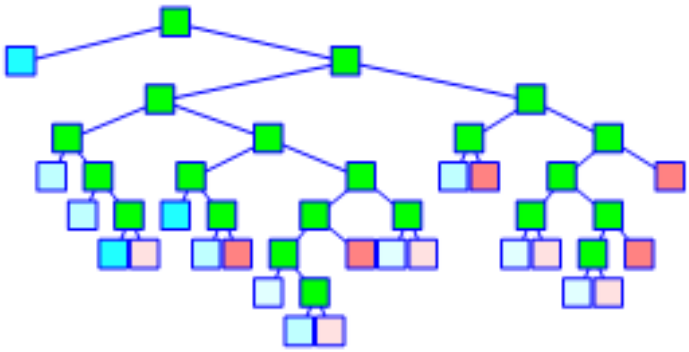

Figure 4: Regression tree topology

Regression trees are used for predicting the membership of cases in the classes of a categorical target variable from their measurements on one or more explanatory variables. The target variable, in this case, is the Living place, which is a categorical variable and takes the values of "1" for "family living in urban" and "0" for "family living in rural". In order to predict the value of the target variable using the regression tree, the model uses the values of the explanatory variables to move through the tree until it reaches a terminal node, and then ultimately predicts the category shown for that node as in Fig. 5 and Fig.6.

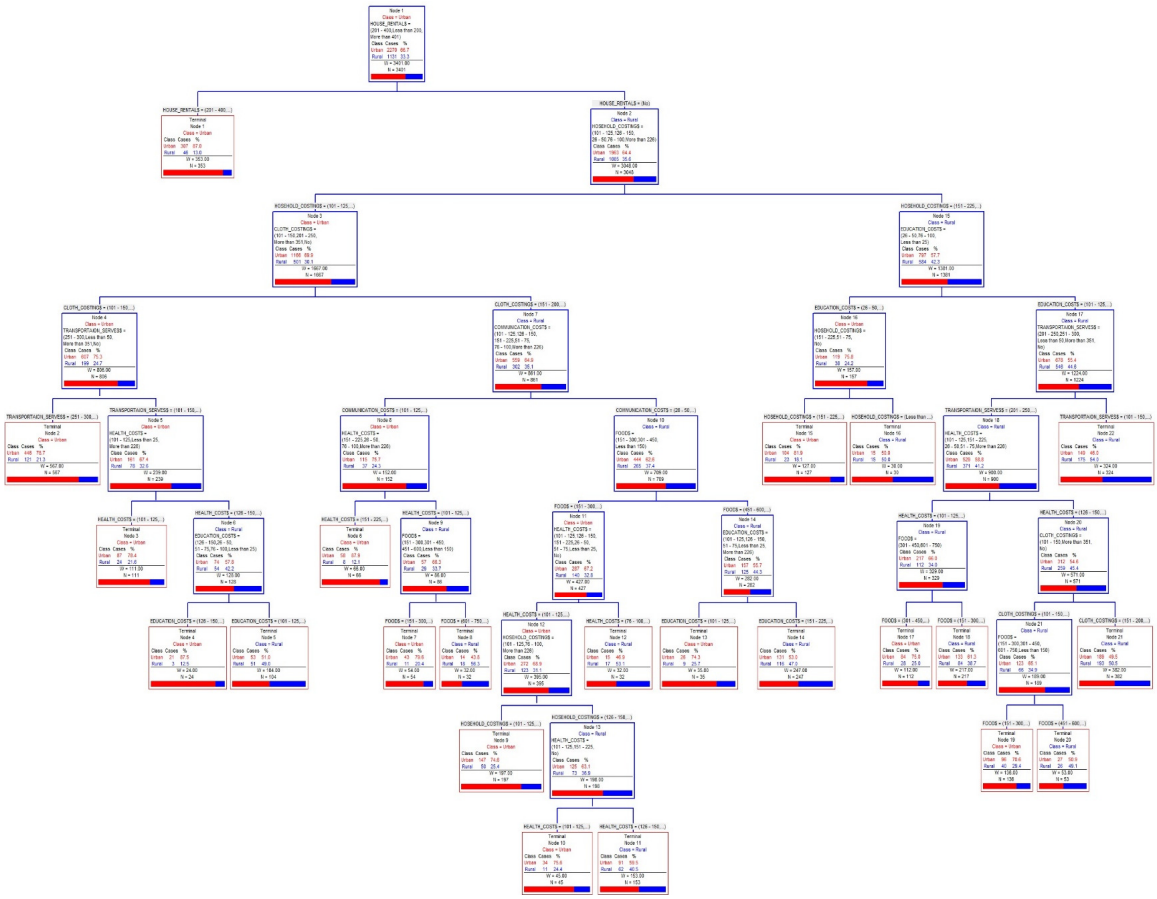

Figure 5: optimal Regression Tree

DOI: http://dx.doi.org/10.25098/3.1.9

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

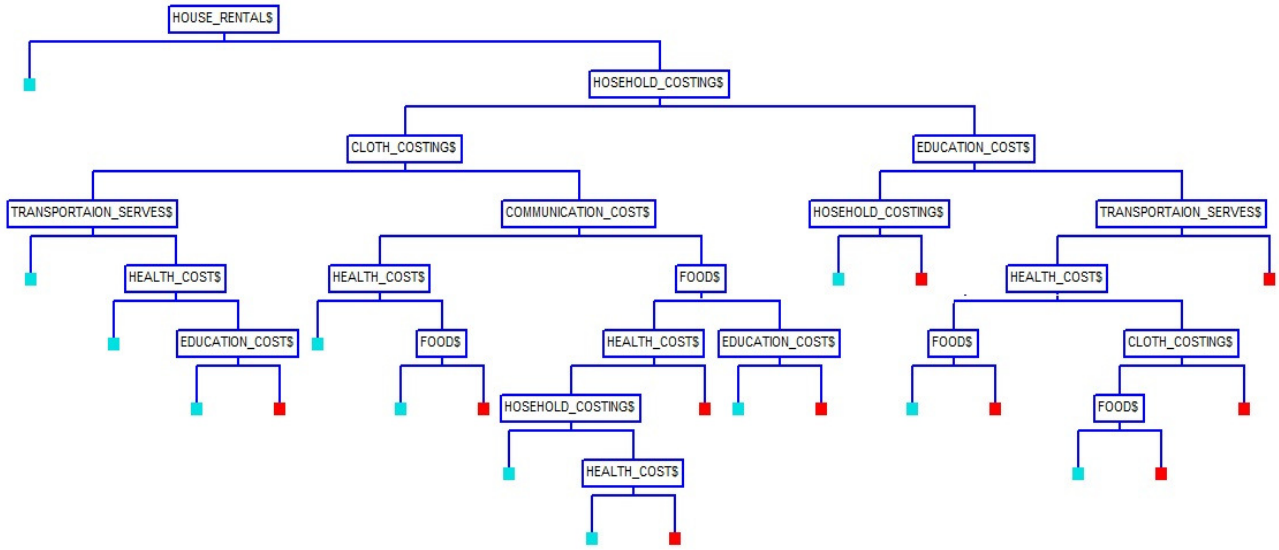

Figure 6: optimal Regression Tree

To explore the tree, it is noted that the splitting criteria from node 2 is house rental (No) which contains a total 1963 of families live in urban, with a split of 1166 families into node 3, and 797 families into node 15, the splitting criteria from node 3 is household costing was ranged (101000 – 125000). The total 1166 of families live in urban, with a split of 607 families into node 4, and 559 families into node 7. From the node 4 the splitting criteria is cloth costing was ranged (101000 – 150000), also the splitting criteria from the node 7 is cloth costing was ranged (151000 – 200000). the total 607 families live in urban, with a split of 161 families into node 5 and 446 families into terminal node 2. From node 5 the splitting criteria is transportation serves was ranged (101000 – 150000). the total 161 families live in urban in the node 5, with a split of 74 families into node 6 and 87 families into terminal node 3. From node 6 the splitting criteria is health costing was ranged (126000 – 150000). The total 74 families in node 6 that live in urban, with a split of 21 families into terminal node 4, and 53 families into terminal node 5. Although the total 559 families live in urban in the node 7, with a split of 115 families into node 8, and 444 families into node 10, the splitting criteria from the node 8 is communication cost was ranged (101000 – 125000). The total 115 of families live in urban, with a split of 58 families into terminal node 6, and 57 families into node 9. From the node 9 the splitting criteria is health costing was ranged (101000 – 150000), the total 57 families live in urban in the node 9, with a split of 43 families into terminal node 7 and 14 families into terminal node 8.

From node 10 the splitting criteria is communication cost was ranged (26000 – 50000), the total 444 families live in urban, with a split of 157 families into node 14 and 287 families into node 11. From node 14 the splitting criteria is food was ranged (451000 – 600000),

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

with a split of 26 families into terminal node 13, and 131 families into terminal node 14. In addition the total 287 families live in urban in the node 11, with a split of 272 families into node 12, and 15 families into terminal node 12, the splitting criteria from the node 12 is health cost was ranged (101000 – 125000). The total 272 of families live in urban, with a split of 147 families into terminal node 9, and 125 families into node 13, where the splitting criteria for the same node (13) is household costing was ranged (126000 – 150000), the total 125 families live in urban in the node 13, with a split of 34 families into terminal node 10 and 91 families into terminal node 11. Next the splitting criteria from the node 15 is household costing was ranged (151000 – 225000). The total 797 of families live in urban, with a split of 119 families into node 16, and 678 families into node 17. From the node 16 the splitting criteria is education cost was ranged (26000 – 50000) is split to two terminal node which are terminal node 15 with 104 families and 15 families in to the terminal node 16, also the splitting criteria from the node 17 is education cost was ranged (101000 – 125000). the total 678 families live in urban, with a split of 529 families into node 18 and 149 families into terminal node 22. From node 18 the splitting criteria is transportation serves was ranged (201000 – 250000). the total 529 families live in urban, with a split of 217 families into node 19 and 312 families into node 20. From node 19 the splitting criteria is health costing was ranged (101000 – 125000) is split into two terminal node which are terminal node 17 with 84 families and 133 families into terminal node 18, while the total 312 families live in urban, with a split of 123 families into node 21 and 189 families into terminal node 21. From node 21 the splitting criteria is cloth costing was ranged (101000 – 150000) is split into two terminal node which are terminal node 19 with 96 families and 27 families into terminal node 20.

3.3 Model assessment

The proposed model shows excellent performance with a test value of the ROC of 0.68 and train value of ROC is 0.70. ROC can range between 0 and 1 with higher values indicating better performance as shown in Table 2 and Fig.7, for this tree, a test accuracy is 59.61%, and the percent of error misclassification rate of class urban is 37.4%, the train accuracy is of 66.93%, and the percent of error misclassification rate of class urban is 35.9%, while the ratio of correct classification 62.5% and 64.01% for testing and training data respectively as in Table 2. [21] The ROC score for each train sample was indicating that the model described the data well. On the other hand, ROC scores on test samples after cross validation was reassuring in prediction on future similar sample [21].

DOI: http://dx.doi.org/10.25098/3.1.9

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Table 2: Prediction Success – Test and Train

Actual Class	Total Class		Percent Correct		Predicted Classes			
					Urban		Rural	
					N = 1827	N = 338	N = 1574	N = 274
	Train	Test	Train	Test	Train	Test	Train	Test
Urban	2,270	409	64.01%	62.59%	1,453	256	817	153
Rural	1,131	203	66.93%	59.61%	374	82	757	121
Total:	3,401	612						
Average:			65.47%	61.10%				
Overall % Correct:			64.98%	61.60%				
Specificity			66.93%	59.61%				
Sensitivity/Recall			64.01%	62.59%				
Precision			79.53%	75.74%				
ROC			70.93%	68.54%				

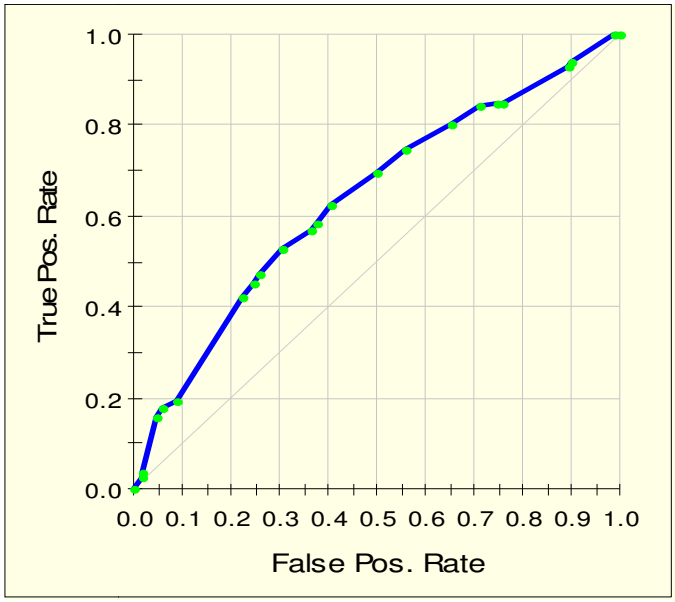

Figure 7: ROC curve Result for train and test data for 23 Nodes

3.4 Variables important

The general criterion for selecting the explanatory variable at each node and its cut point value is the minimum deviation from all possible explanatory s and threshold values. Defining a given node as a terminal one based on the minimum error achieved as per a preset criterion for the minimum number of observations or some other type of restriction [15]. In this study can be determine the relative importance of each variable within the construction of the tree, which is given as house rental (100.00%), household costing (94.19%), education coast (84.28%), health cost (63.26%), cloth costing (57.27%), food

The Scientific Journal of Cihan University – Sulaimanyia **PP: 130-150**
Volume (3), Issue (1), Jun 2019
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

(54.53%), communication cost (47.28%), and transportation serves variable has less important contribution in the study which equal to (44.78%) as shown in Table 3.

Table 3: Variable Importance in the study

Variable	Score
HOUSE_RENTAL	100.00
HOSEHOLD_COSTING	94.19
EDUCATION_COST	84.28
HEALTH_COST	63.26
CLOTH_COSTING	57.27
FOOD	54.53
COMMUNICATION_COST	47.28
TRANSPORTAION_SERVES	44.78

4. Results and conclusions

The results of the study are shown form 307 families who live in urban expend (201000 – 400000) dinar of their monthly expenditures to house rent, from 797 families who live in urban area in Sulaimani Governorate, the (151000 - 225000) dinar of their monthly expenditures expend to household. Also, 678 of these families, expend (101000 – 125000) dinar to education serves and 149 families of them expend (101000 – 150000) of their monthly expenditure to transportation serves. In the node 16, 119 families who live in Sulaimani city expend (26000 – 50000) dinars of their monthly expenditures to education serves. From these families 104 families expend (151000 – 225000) dinar to household. While 15 of these families less than 25000 dinars expend to household. In node 18, 529 families who live in Sulaimani city expend (201000 – 250000) dinars to transportation serves which 217 families of them expend (101000 – 125000) dinars to health, also 84 families of them expend (301000 – 450000) dinars of their expenditures to food, at the same time 133 families of them expend (151000 – 300000) dinars to food. While 529 families who live in Sulaimani city in node 18, with a split of 312 families in the node 20, who spend (126000 – 150000) dinar to health field, 123 families of them expend (101000 – 150000) dinars to cloth coasting, while 96 families of these 123 families expend (151000 – 300000) dinars to food. In the other hand 27 families of these 123 who live urban area in Sulaimani city expend (451000 – 600000) dinars to food, also 189 families of 312 families in the node 20 expend (151000 – 200000) to cloth costing. In the node 3, from 1166 families who live in urban area in Sulaimani governorate, the (101000 - 125000) dinar of their monthly expenditures expend to household. Also, 559 of these families, expend (151000 – 200000) dinar to cloth costing, and 444 families of them expend (101000 – 150000) of their monthly expenditure to communication serves, while 157 families of them expend (451000 – 600000) of their monthly expenditure to food, then from these 157 families 26 families expend (101000 – 125000) to education serves and 131 families expend (151000 – 225000) to education serves. In the node 11, 287 families who live in

DOI: http://dx.doi.org/10.25098/3.1.9

Sulaimani city expend (151000 – 300000) dinars of their monthly expenditures to food, from these families 272 families expend (101000 – 125000) dinar to health serves, while 125 of these families expend (126000 – 150000) dinars to household and from these families 91 families expend (126000 – 150000) dinars for health serves. While in the terminal nodes 12, 9, and 10, there are 15, 147, and 34 families who expend (76000 – 100000), (101000 – 125000), and (101000 – 125000) dinars to health serves, household, and health serves respectively. In the node 8, 115 families who live in Sulaimani city expend (101000 – 125000) dinars of their monthly expenditures to communication serves, from these families 57 families expend (101000 – 125000) dinar to health serves, while 14 of these families expend (601000 – 750000) dinars to food, and in the terminal nodes 6, and 7, there are 58, and 43 families who expend (151000 – 225000), and (151000 – 300000) dinars to health serves, and food respectively.

Finally in the node 4, 607 families who live in Sulaimani city expend (101000 – 150000) dinars of their monthly expenditures to cloth, from these families 161 families expend (101000 – 150000) dinar to transportation serves, while 74 of these families expend (126000 – 150000) dinars to health and from these families 53 families expend (101000 – 125000) dinars for education serves. While in the terminal nodes 2, 3, and 4, there are 446, 87, and 21 families who expend (251000 – 300000), (101000 – 125000), and (126000 – 150000) dinars to transportation serves, health, and education serves respectively.

5. References

1. Abdul Kareem S., Raviraja S., Awadh N. A., Kamaruzaman A. & Kajindran A. (2010) "Classification and Regression Tree In Prediction Of Survival Of AIDS Patients" Malaysian Journal of Computer Science, Vol. 23(3), 2010.
2. Bae H., Olson B. H., Hsu K. & Sorooshian S. (2010) "Classification and regression tree (CART) analysis for indicator bacterial concentration prediction for a Californian coastal area" Water Science & Technology—WST | 61.2 | 2010, doi: 10.2166/wst.2010.842.
3. Bevilacqua, M., Braglia, M., Montanari, R., 2003. The classification and regression tree approach to pump failure rate analysis. Reliab. Eng. Syst. Saf. 79 (1), 59–67.
4. Cascini, L., Ciurleo,M., Di Nocera, S., Gulla, G., 2015. A new-old approach for shallow landslide analysis and susceptibility zoning in fine-grained weathered soils of southern Italy. Geomorphology 241, 371–381.
5. Cheng Z. MD MPH, Nakatsugawa M. Phd, Hu Ch. Phd, Robertson S. P. Phd, Hui X. MD MS, Moore J. A. Phd, Bowers M. R. BS, Kiess A. P. MD Phd, Page B. R. MD, Burns L. BSN, Muse M. BSN, Choflet A. MS RN OCN, Sakaue K. MS, Sugiyama Sh. MS, Utsunomiya K. MS, Wong J. W. Phd, McNutt T. R. Phd and Quon H. MD MS (2018) "Evaluation of classification and regression tree (CART) model in weight loss prediction following head and neck cancer radiation therapy" Advances in Radiation Oncology (2018) 3, 346–355. https://doi.org/10.1016/j.adro.2017.11.006.

6. Conoscenti, C., Ciaccio, M., Caraballo-Arias, N.A., Gomez-Gutierrez, A., Rotigliano, E., Agnesi, V., 2015. Assessment of susceptibility to earth-flow landslide using logistic regression and multivariate adaptive regression splines: a case of the Bence River basin (western Sicily, Italy). Geomorphology 242, 49–64.
7. D. Steinberg, P.L. Colla, CART: Tree-structured nonparametric data analysis, Salford Systems, San Diego, CA, 1995.
8. Daniels S., and Glorieux I. (2015) "Convenience, food and family lives. A socio-typological study of household food expenditures in 21st-century Belgium" Appetite. Volume 94, 1 November 2015, Pages 54-61. https://doi.org/10.1016/j.appet.2015.04.074.
9. Dehnavi, A., Aghdam, I.N., Pradhan, B., Varzandeh, M.H.M., 2015. A new hybrid model using step-wise weight assessment ratio analysis (SWAM) technique and adaptive neuro-fuzzy inference system (ANFIS) for regional landslide hazard assessment in Iran. Catena 135, 122–148.
10. Felicísimo, Á.M., Cuartero, A., Remondo, J., Quirós, E., 2013. Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. Landslides 10 (2), 175–189.
11. Guo, C.B., Montgomery, D.R., Zhang, Y.S., Wang, K., Yang, Z.H., 2015. Quantitative assessment of landslide susceptibility along the Xianshuihe fault zone, Tibetan plateau, China. Geomorphology 248, 93–110.
12. Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression. A Wiley-Interscience Publication, New York.
13. Hussin, H.Y., Zumpano, V., Reichenbach, P., Sterlacchini, S., Micu, M., van Westen, C., Balteanu, D., 2016. Different landslide sampling strategies in a grid-based bi-variate statistical susceptibility model. Geomorphology 253, 508–523.
14. Iliev I. P., Voynikova D. S., & Gocheva-Ilieva S. G. (2013) "Application of the Classification and Regression Trees for Modeling the Laser Output Power of a Copper Bromide Vapor Laser" Hindawi Publishing Corporation http://dx.doi.org/10.1155/2013/654845.
15. Josephine W. Mburu, Leonard Kingwara, Magiri Ester, Nyerere Andrew, 2018." Use of classification and regression tree (CART), to identify hemoglobin A1C (HbA1C) cut-off thresholds predictive of poor tuberculosis treatment outcomes and associated risk factors", J Clin Tuberc Other Mycobact Dis, https://doi.org/10.1016/j.jctube.2018.01.002, journal homepage: www.elsevier.com/locate/jctube.
16. Koon, S., Petscher, Y., 2015. Comparing Methodologies for Developing an Early Warning System: Classification and Regression Tree Model versus Logistic Regression. REL 2015–077. Southeast, Regional Educational Laboratory.

DOI: http://dx.doi.org/10.25098/3.1.9

17. Li Yong, 2006. "Predicting materials properties and behavior using classification and regression trees", Materials Science and Engineering A 433, 261–268, doi:10.1016/j.msea.2006.06.100.
18. Mahjoobi, J., Etemad-Shahidi, A., 2008. An alternative approach for the prediction of significant wave heights based on classification and regression trees. Appl. Ocean Res. 30 (3), 172–177.
19. Malinowska, A., 2014. Classification and regression tree theory application for assessment of building damage caused by surface deformation. Nat. Hazards 73 (2), 317–334.
20. Park T. (2018) "Health care utilization and expenditures among adults with rheumatoid arthritis using specialty pharmaceuticals" Research in Social and Administrative Pharmacy. https://doi.org/10.1016/j.sapharm.2018.09.003.
21. Pasipanodya JG, Gumbo T. (2011). A new evolutionary and pharmacokinetic pharmacodynamic scenario for rapid emergence of resistance to single and multiple antituberculosis drugs. Curr Opin Pharmacol;11:457–63.
22. Pouliakis A., Karakitsou E., Chrelias Ch., Pappas A., Panayiotides I., Valasoulis G., Kyrgiou M., Paraskevaidis E. and Karakitsos P. (2015) "The Application of Classification and Regression Trees for the Triage of Women for Referral to Colposcopy and the Estimation of Risk for Cervical Intraepithelial Neoplasia" Hindawi Publishing Corporation, http://dx.doi.org/10.1155/2015/914740.
23. Roger, J. (2000). An introduction to classification and regression tree (CART) analysis, Emergency Medicine Dep., Torrance, California Annual Meeting of the society for academic emergency medicine in San Francisco.
24. Salari M. and Javid R. J. (2017) "Modeling household energy expenditure in the United States" Renewable and Sustainable Energy Reviews. Volume 69, March 2017, Pages 822-832. https://doi.org/10.1016/j.rser.2016.11.183.
25. sampurno, Fanny. (2006). identifying risk factors associated with new onset cardiovascular disease in patients with type I diabetes using classification tree, Melbourne University, Honours Thesis, Mathematics and statistics Dep.
26. Sarti S., Terraneo M. and Bordogna M. T., (2017) "Poverty and private health expenditures in Italian households during the recent crisis" Health Policy. Volume 121, Issue 3, March 2017, Pages 307-314. https://doi.org/10.1016/j.healthpol.2016.12.008.
27. Soo Yeon Choi, Il Won Seo,2018" Prediction of Fecal Coliform using Logistic Regression and Tree-based Classification, Journal of Hydro-environment Research Volume 21, October 2018, Pages 96-108, https://doi.org/10.1016/j.jher.2018.09.002.
28. Tien Bui, D., Lofman, O., Revhaug, I., Dick, O., 2011. Landslide susceptibility analysis in the Hoa Binh province of Vietnamusing statistical index and logistic regression. Nat. Hazards 59 (3), 1413–1444.

29. Tien Bui, D., Nguyen, Q.-P., Hoang, N.-D., Klempe, H., 2016a. A novel fuzzy K-nearest neighbor inference model with differential evolution for spatial prediction of rainfall- induced shallow landslides in a tropical hilly area using GIS. Landslides http:// dx.doi.org/10.1007/s10346-016-0708-4.
30. Wang, L.-J., Guo, M., Sawada, K., Lin, J., Zhang, J., 2015a. Landslide susceptibility mapping in Mizunami City, Japan: a comparison between logistic regression, bivariate statistical analysis and multivariate adaptive regression spline models. Catena 135, 271–282.
31. Yang, T., Gao, X., Sorooshian, S., Li, X., 2016. Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. Water Resour. Res. 52 (3), 1626–1651.
32. Yoo K., Yoo H., Lee J. M., Shukla S. K. & Park J. (2018) "Classification and Regression Tree Approach for Prediction of Potential Hazards of Urban Airborne Bacteria during Asian Dust Events" Scientific Reports | (2018) 8:11823 | DOI: 10.1038/s41598-018-29796-7.
33. Zhao Yannan, Li Yuan, Zhang Lifen, Wang Qiuliang, 2016," Groundwater level prediction of landslide based on classification and regression tree", geodesy and geodynamics, vol. 7, no.5, 348-355, http://dx.doi.org/10.1016/j.geog.2016.07.005.
34. Zimmerman R. K., Balasubramani G. K., Nowalk M. P., Eng H., Urbanski L., Jackson M. L., Jackson L. A., McLean H. Q., Belongia E. A., Monto A. S., . Malosh R. E., Gaglani M., Clipper L., Flannery B. and Wisniewsk S. R. (2016) "Classification and Regression Tree (CART) analysis to predict influenza in primary care patients" Zimmerman et al. BMC Infectious Diseases (2016) 16:503 https://doi.org/10.1186/s12879-016-1839-x.

DOI: http://dx.doi.org/10.25098/3.1.9