Middle East Journal for Scientific Publishing

Vol. 4, Issue No. 1, 80-94 (2021)

E-ISSN: 2707-188X

Predicting Education Dropout in Benghazi City and its suburbs by Using Classification Trees

Osama H. Othman[1]

E-mail: serser11@yahoo.com

Eyman Musa Farag Farag[2]

E-mail: Eman24_11_1982@yahoo.com

Received: 17 Oct. 2020, Revised: 20 . 2020, Accepted: 22 Nov. 2020

Published online: 6 Feb. 2021

Abstract

The education sector is considered to be one of the most important sectors in any society , it is the corner stone toward the progress and development , therefore ministers of educations, presidents and heads of universities and schools have prepared the education cadres to enable them to participate in building the society. Despite the attention that countries pay toward education, still there is a big problem facing this attention that most countries have failed to solve, it is the dropout of education. In this work our aim is the prediction of the factors (variables) that likely influence education dropout. The utilized statistical model for that purpose is classification trees, which is a nonparametric data mining device that suit our

[1] Lecturer, Department of Statistics, Ajdabiya University, Libya.
[2] Lecturer, Department of Statistics, Ajdabiya University, Libya.

database and meet our aim of the study. The target here is to predict students who will probable dropout the study at secondary school and not proceed their study to universities or high institutes.

Keywords: Data mining , Classification trees , Multivariate Analysis , Prediction

1.0 Introduction

We may determine that the concepts of dropout of education as a serious problem faces the education sector in various societies, it creates a large bleeding in depriving many of those who are willing to learn, because the dropouts are considered as a burden on those countries that negates their educational and technological progress, opinions were varied in determining the concept of dropout of education , In 1992 the UNICEF believed that the dropout of education is that when the children do not go to school or dropout before successfully completing that stage of schooling whether willingly or due to other circumstances , also the lack of persistence in attending school for one year or more.

Allaqani in (1999) believed that the dropout is anyone who dropout from school at any level regardless of the degree of the level. Almahanna (2001) believed that the dropout of education when the pupils totally stop attending school and dropout after joining it , whether this stoppage happened immediately after joining school or after that by one or more of school years before completing that fixed stage.

Whereas Bukour in (2003) believed that the dropout student is the one who dropout before the end of school year. Linda in (2006) believed that the dropout is the student who was registered in the same date of the current year as that of last year . or the one who was not registered in the beginning of the previous school year and was expected to register in the current year, but he did not during the year. Or the one who did not graduate from Junior High School , or the student who did not transferred to another school and was absent because of illness or death. Some

believed that the dropout is the person who left the school and never return to it or to another school , whereas the person who never attended school is not considered dropout but illiterate. Any child started a certain school stage and did not successfully complete it, is considered on a case of dropout.

After all our scope in this study on the factors which effect mostly in dropping out the study of the students at the high or secondary school. These factors which prevents the students to proceed their academic studies at universities , colleges and institutes is of our interest in order to study them in further for the purpose of eliminate them or at least to decrease their influence on the society.

2.0 Methodology

A number of suggested statistical techniques might be used for our consideration, but the most appropriate technique that meets our aim and the nature of the collected data is Classification Trees (CT). The Classification Trees is a non parametric multivariate technique which used for the purposes of classification and prediction. Being a non parametric method that's make it free from many assumptions that is needed in case of parametric methods.

Classification trees models have many usages in a wide Varity of disciplines including : Variable selection, assessing the relative importance of variables, Prediction, data manipulation, Classification, segmentation, stratification, data reduction and many others usages. Our scope is pointed out to prediction that is based on classification.

*** How to work?**

The prediction is made by creating a tree based classification model. The model works by classifying cases into well defined groups, that is. The predicted values of the dependent (target) variable can be obtained on the basis of the independent (predictors) variables values.

That process which entails recursive portioning has many procedures for checking the quality of partition, model prediction accuracy, Pruning, validity as well as the growing method. Some of the previous procedures will be checked later and some were done by default of SPSS. Classification trees procedure offers many growing methods such as CHAID, exhaustive CHAID, CRT and QUEST.

The validation is proceed in Classification Trees in order to check its performance. The cross validation works by dividing the whole data set into subsamples (folds), the tree is generated as many as the number of selected folds. The first tree is based on all of the cases except those in the first sample fold, the second tree is based on all of the cases except those in the second sample fold, and so on.

The misclassification risk is estimated for each subsample (fold) individually by applying the tree on all cases except the sample fold excluded. Evaluating the final tree model can be done by the misclassification risk that is calculated by averaging the risks for all trees.

3.0 Data , Application and Results

The data were collected through a sample size of 1657 through a questionnaire that was designed to include all the variables that are suspected in influencing the phenomenon under study. The sample was taken randomly within the community in the city of Benghazi and its suburbs by simple random sampling.

The dependent variable and the explanatory variables in the study are defined as follows:-

Table (1): Description of variables

Variable's	Variable's Title	Possible Values
(Y)	Class dropout from school.	Illiterate , Elementary , Preparatory , Secondary
X1	Age	

X2	Gender	Male , Female
X3	Number of family members	
X4	Order of researched in the family	
X5	Father's occupation	Employee , Unemployed
X6	Age at drop of the study	
X7	Cause of dropout is The school	Yes , No
X8	Cause of dropout The teacher	Yes , No
X9	Cause of dropout is Superintendent	Yes , No
X10	Cause of dropout is Separation of parents	Yes , No
X11	Cause of dropout is Family economic status	Yes , No
X12	Cause of dropout is Death of father	Yes , No
X13	Cause of dropout is Death of mother	Yes , No
X14	Cause of dropout is Health causes	Yes , No
X15	Cause of dropout is other circumstances	Yes , No
X16	Occupation of the research subject	Unemployed , Technical Occupation , Management , Services , Odd Jobs
X17	Father's age	
X18	Mother's age	
X19	Father's education level	Illiterate , Primary , Preparatory , Secondary, Academic
X20	Mother's education level	Illiterate , Primary , Preparatory , Secondary, Academic
X21	Mother's occupation	Employee , Unemployed
X22	The family monthly income	Less Than 250 , 500 - 250 , Over 500
X23	Housing status	Proper , Improper

The independent (predictors) variables displayed in table() are mixed of quantitative (numeric) and qualitative (categorical), this variety in the data suit the usage of classification trees. The dependent (response) variable as it was shown previously is qualitative as it assumed and required by Classification Trees.

The application was done by the statistical software SPSS version 23. The steps of the application are presented with the explanations in ordered way as follows:-

All the growing methods were applied to find the best generated tree, the methods which were applied CHAID, Exhaustive CHAID and CRT. The Quest method was excluded because it needs a nominal dependent variable where the dependent variable here is ordinal. Many comparisons were made between the three used growing methods to find the best generated tree.

3.1 Evaluating the model

The model results many charts and tables that used in evaluating it in many aspects. The risk table provides a brief evaluation of the quality of the model for all the three used growing methods by resubstitution and Cross Validation methods. The resubstitution estimates shows that the CHAID methods gave less estimated risk than CRT method, whereas the Cross Validation estimates show that the three methods have *approximately* the same estimated risk. The standard error values is the same for all methods. In short, the tree growing methods are almost the same.

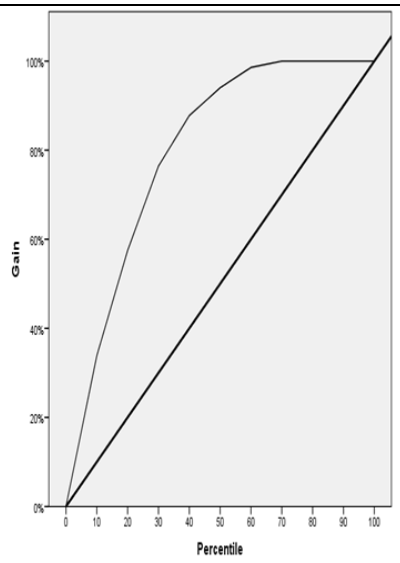
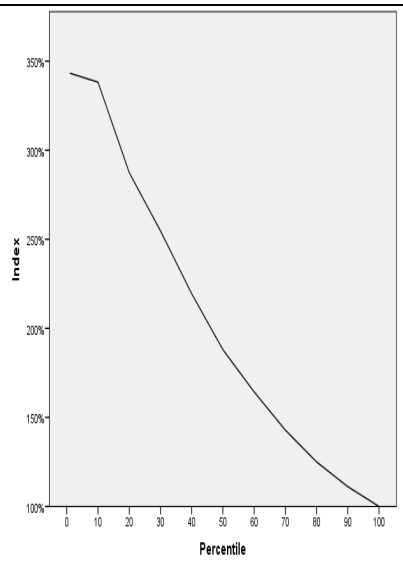
Table(2) : Risk estimation

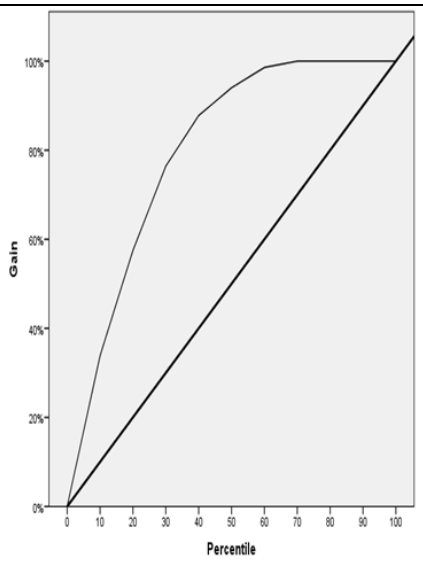
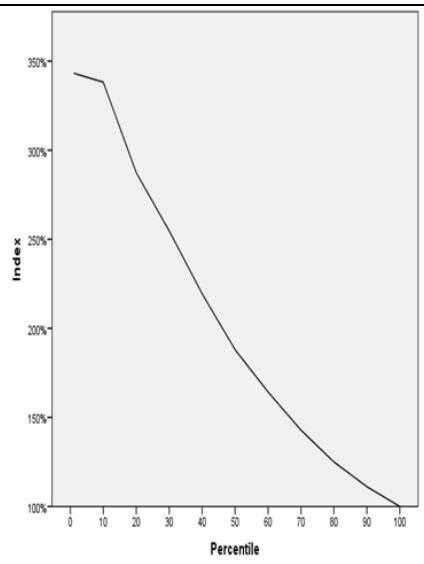
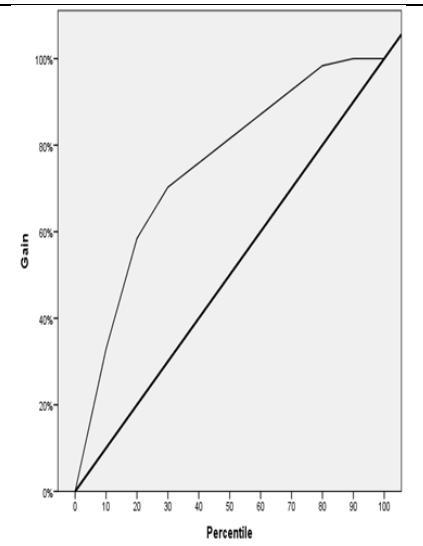
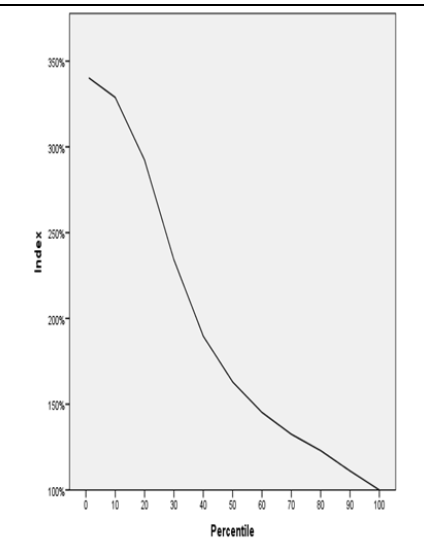
Growing Method	Method	Estimate	Standard Error
CHAID	Resubstitution	0.318	0.012
	Cross Validation	0.334	0.012
Exhaustive CHAID	Resubstitution	0.318	0.012
	Cross Validation	0.328	0.012
CRT	Resubstitution	0.324	0.012
	Cross Validation	0.331	0.012

Table (3) presents the overall percent correct for the three growing methods. The CHAID methods give 68.0% which is higher correct than CRT method 67.6%.

Table(3) : Classification overall percent correct	
Growing Method	Overall percent correct
CHAID	68.2%
Exhaustive CHAID	68.2%
CRT	68.0%

The cumulative Gain and Index charts are presented in graph(1). For all of growing methods it can be seen that the models are good. All gain charts don't follow the diagonal line, which means the model provides well information. The cumulative indices in each chart start above 100% and decline regularly until get to 100% and that is a good indicator of the model adequacy.

Growing Method	**Gain Chart**	**Index Chart**
CHAID		

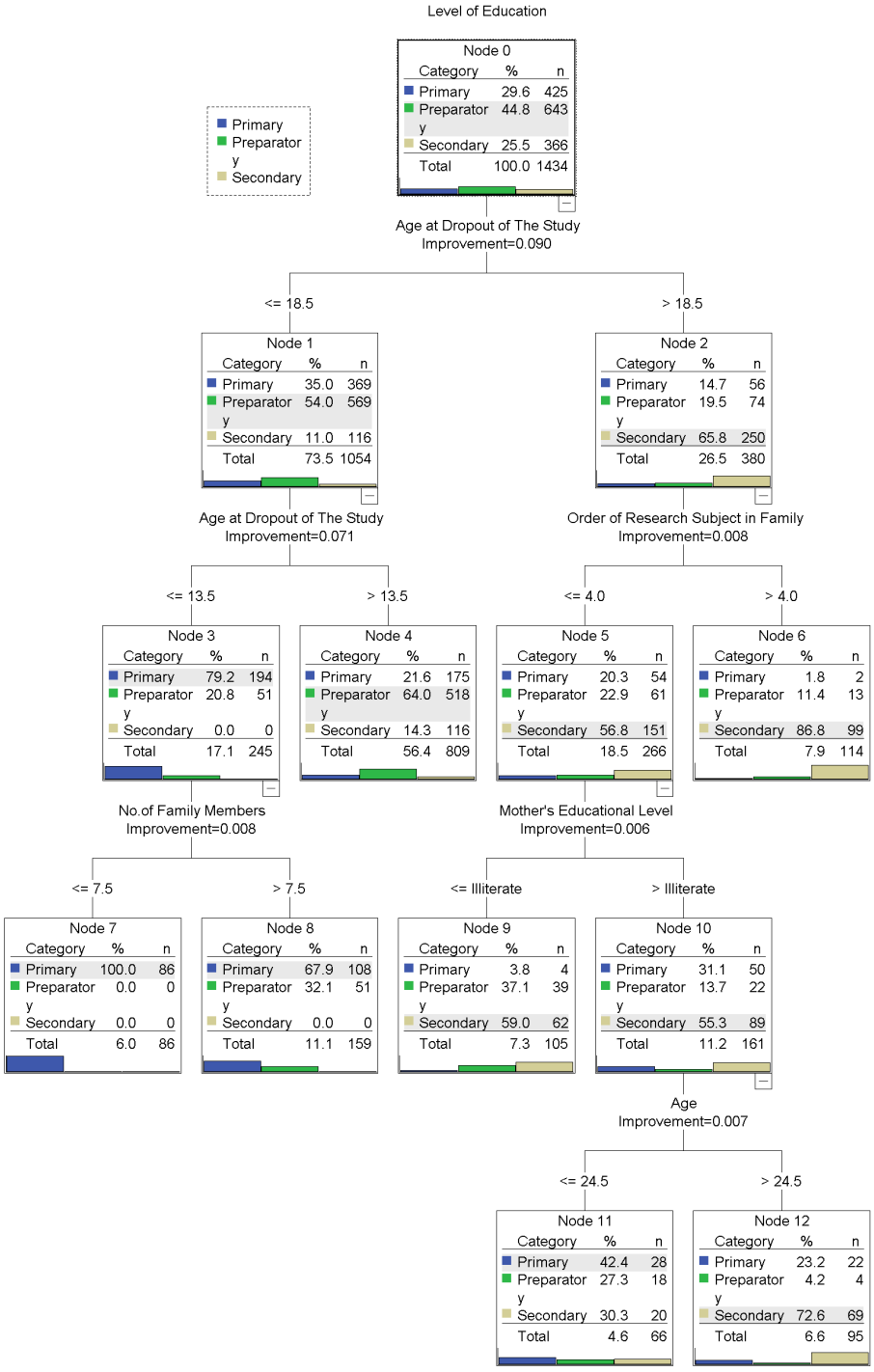
Exhaustive CHAID		
CRT		

Graph (1): Gain and Index charts for the three used growing methods

After presenting these calculations about risk and correct percent of classification which show a large resemblance between growing methods. the rules of classifications trees for the three growing methods for terminal nodes are shown with the tree graph for each method.

The focus will be just on terminal nodes that belong to the target category (Secondary) and describe the classification rules which show the students who will likely drop the study out at this stage.

*** CRT Method**



Graph(2): Generated tree by CRT method

Node 6 : Student will dropout the study at secondary school with probability 0.868

If the Age at drop out of the study is greater than 18.5 and Order of Research Subject in Family is greater than 4 .

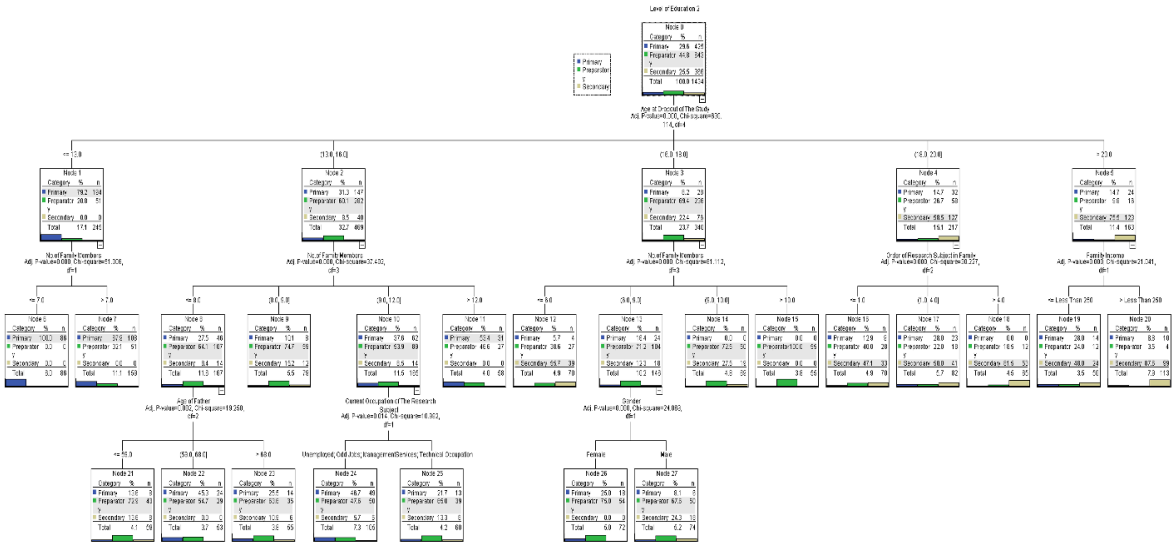
Node 9 : Student will dropout the study at secondary school with probability 0.59

If the Age at drop out of the study is greater than 18.5 and the order of research subject in family is less than 4 and the mother's educational level is illiterate or lower than that.

Node 12 : Student will dropout the study at secondary school with probability 0.726

If the Age at drop out of the study is greater than 18.5 and the order of research subject in family is less than 4 and the mother's educational level is illiterate or higher than that or and the age is larger than 24.5 .

CHAID Method :



Graph (3) : Generated tree by CHAID method

Node 12 : Student will dropout the study at secondary school with probability 0.557

If the age at drop out of the study is between (16 - 18) and the number of family members is less than or equal to 6.

Node 16 : Student will dropout the study at secondary school with probability 0.471

If the age at drop out of the study is between (18 - 20) and the Order of Research Subject in Family is the first.

Node 17 : Student will dropout the study at secondary school with probability 0.50

If the age at the drop out of the study is between (18 - 20) and the order of research subject in family is between (1 - 4) .

Node 18 : Student will dropout the study at secondary school with probability 0.815

If the age at the drop out of the study is between (18 - 20) and the order of research subject in family is higher than the fourth .

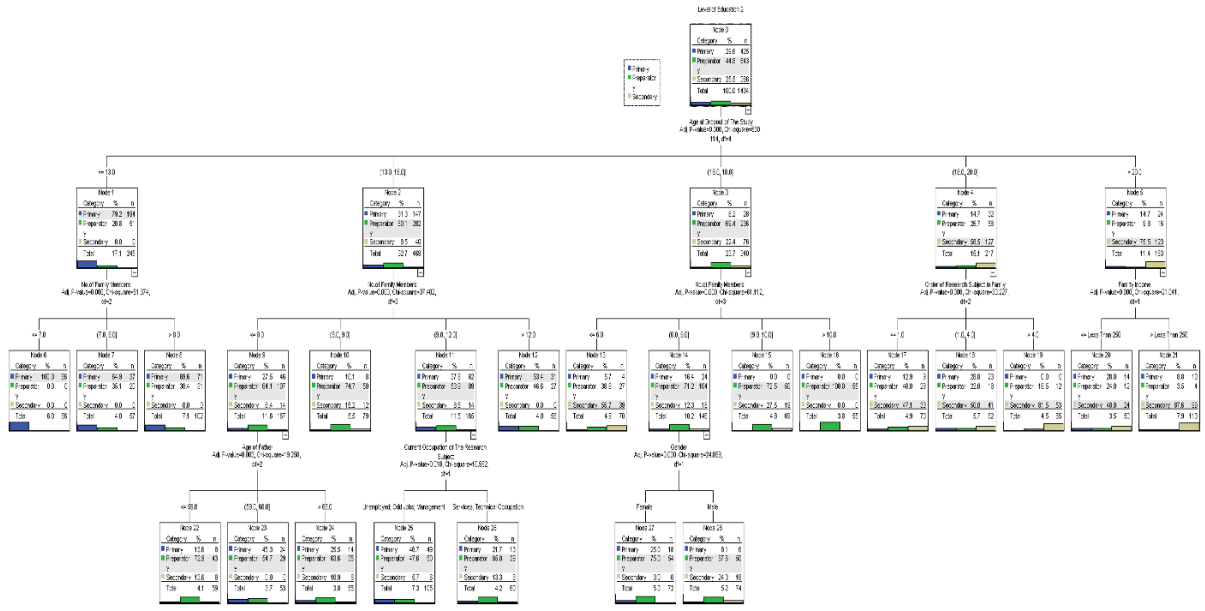
Node 19 : Student will dropout the study at secondary school with probability 0.48

If the age at the drop out of the study is larger than 20 and the family income is less than or equal to 250 .

Node 20 : Student will dropout the study at secondary school with probability 0.876

If the age at the drop out of the study is larger than 20 and the family income is less than 250 .

*** Exhaustive CHAID**



Graph (4) : Generated tree by Exhaustive CHAID method Node 13 : Student will dropout the study at secondary school with probability 0.557

If the age at the drop out of the study is between (16 - 18) and the number of family members is less than or equal to 6.

Node 17 : Student will dropout the study at secondary school with probability 0.471

If the age at the drop out of the study is between (18 - 20) and the order of research subject in family is the first.

Node 18 : Student will dropout the study at secondary school with probability 0.50

If the age at the drop out of the study is between (18 - 20) and the order of research subject in family is between (1 - 4).

Node 19 : Student will dropout the study at secondary school with probability 0.815

If the age at the drop out of the study is between (18 - 20) and the order of research subject in family is greater than 4.

Node 20 : Student will dropout the study at secondary school with probability 0.48

If the age at the drop out of the study is larger than 20 and the family income is less than or equal to 250.

Node 21 : Student will dropout the study at secondary school with probability 0.876

If the age at the drop out of the study is larger than 20 and the family income is less than 250.

4.0 Conclusion and Recommendations

The study aims met with the used technique successfully. Also the model evaluation procedures showed that the models are valid and their predictions will be almost correct. The three growing methods were semi similar in the evaluation measures and graphs and for that reason their results were all presented, in addition the classification rules were approximately the same and there is a resemblance in the predictor variables used in splitting.

It can be seen that the most important variables for the three growing methods are : Age at drop out of the study ,Order of Research Subject in Family ,mother's educational level, age ,number of family members and family income .The differences between the shapes of tree growing method were small, where the simplest is the tree of CRT method.

Since All of the growing methods fit the data adequately. Then the differentiation between them would be made in terms of the simplicity and about the values of prediction probabilities.

As for the model simplicity, the CRT growing tree was the simplest and the highest probabilities of prediction was with this method. The trees of CHAID and

Exhaustive CHIAD are more complex than CRT and they have lower probabilities of prediction.

Finally, our conclusion is that choosing any of the three growing methods would not make any difference or effect the prediction in this study. Further investigations can be done by extend the application of Classification Tress to get benefits of the predicted values or by applying a new device in the context of non-parametric multivariate techniques.

References

Bayer, J., Bydzovska, H., Geryk, J., Obsivac, T., Popelinsky, L.(2012). *Predicting drop-out from social behaviour of student*s: Paper presented at the International Conference on Educational Data Mining (EDM) (5th, Chania, Greece, Jun 19-21, 2012)

Bharadwaj,K,B., Pal, S .(2012).*"Mining Educational Data to Reduce Dropout Rates of Engineering Students*". International journal of multidisciplinary sciences and engineering, vol. 3, no. 5, May 2012, 35-39.

Dekker,G. Pechenizkiy,M.Vleeshouwers, J M. (2009).*"Predicting Students Drop Out*: A Case Study, Eindhoven University".

Jadrić, M. Garača, Ž. Čukušić, M.(2010).*" Student Dropout Analysis with Application of Data Mining Methods*" .Management journal of contemporary management issues UDC 65.012.34:378

Kovaciˇ c, Zlatko J.(2010).*"Early prediction of student success: ´ mining students enrolment data*". In Proceedings of Informing Science & IT Education Conference (InSITE), pp. 647–665. Citeseer, 2010.

Lovenoor, A. Nishant,V. Joshua,B. Jevin,W.(2016). *"Predicting Student Dropout in Higher Education*", DataLab, The Information School, University of Washington, Seattle, WA 98195, USA

Quadri,M. Kalyankar,N.(2010). *"Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques*". Global Journal of Computer Science and Technology, Vol. 10 Issue 2 (Ver 1.0), April 2010.

Rai, S.(2014). *"Student's Dropout Risk Assessment in Undergraduate Course at Residential University*" .International Journal of Computer Applications (0975 – 8887) Volume 84 – No 14, December 2013

Rai, S, Saini,P,. Jain,A,K .(2014). *"Model for Prediction of Dropout Student Using ID3 Decision Tree Algorithm*". International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014): ISSN : 2347 - 8446 (Online).

Subitha,S, Sivakumar ,V, Rajalakshmi,S.(2016). *"Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree*" .Indian Journal of Science and Technology, Vol 9(4), DOI: 10.17485/ijst/2016/v9i4/87032, January 2016: ISSN (Online) : 0974-5645.

Veitch, W.R. (2004). *"Identifying characteristics of high school dropouts":* Data mining with a decision tree model, Presented at the Annual Meeting of the American Educational research Association held on April at San Diego, CA.

Villwock, R., Appio, A., & Andreta, A. A. (2015). *Educational Data Mining with Focus on Dropout Rates*. International Journal of Computer Science and Network Security, 15(3), 17–23.