Comparing Machine Learning Techniques and Classical Approach for child's Education and Alternative Activities in case of Pakistan

By

Namal Kinza

Registration No: PIDE2016FMPHILETS13

Supervised By

Dr. Hafsa Hina

Department of Econometric and Statistics

Pakistan Institute of Development Economics

Islamabad Pakistan

2019

Pakistan Institute of Development Economics

CERTIFICATE

This is to certify that this thesis entitled: **"Comparing Machine Learning Techniques and Classical Approach for Child's Education and Alternative Activities in Case of Pakistan"** submitted by Ms. Namal Kinza is accepted in its present form by the Department of Econometrics and Statistics, Pakistan Institute of Development Economics (PIDE), Islamabad as satisfying the requirements for partial fulfillment of the degree in **Master of Philosophy in Econometrics.**

Supervisor:

Dr. Hafsa Hina
Assistant Professor
PIDE, Islamabad.

External Examiner:

27.5.19

Dr. Zahid Asghar
Professor
School of Economics
Quaid-i-Azam University, Islamabad

Head,
Department of Econometrics and Statistics:

Dr. Amena Urooj

Declaration

I **Namal Kinza** hereby state that my MPhil thesis titled "**Comparing machine learning techniques and classical approach for child's Education and alternative activity in case of Pakistan"** is my own work and has not been submitted previously by me for taking any degree from Pakistan Institute of Development Economics or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my Graduation the university has the right to withdraw my MPhil degree.

Date:___________________ Signature of Student

Dedicated to my Beloved Brother

Umair Ansar

ACKNOWLEDGEMENT

In the name of Allah the most Merciful and Beneficent

First and Foremost praise is to ALLAH, the Almighty, the greatest of all, on whom ultimately we depend for sustenance and guidance. I would like to thank Almighty Allah for giving me opportunity, determination and strength to do my research. His continuous grace and mercy was with me throughout my life and ever more during the tenure of my research.

I would like to express my sincere gratitude to my advisor Dr. Hafsa Hina for the continuous support of my research, for his patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my MPhil study.

I sincerely thank to my Father for financial support during my whole career , special thanks to my brother and my aunt for their encouragement, moral support, personal attention and care .I would like to acknowledge many people for helping me during my research.

Last but not the least, I would like to thank my friends (Hifza ,Lubna ,Rabail and Fatima) for giving valuable suggestions during my study. I enjoyed spending time with them. Thanks for giving me such a joyful time.

Namal Kinza

TABLE OF CONTENTS

LIST OF TABLE vii

LIST OF FIGURE viii

ABSTRACT... ix

CHAPTER 1 1

INTRODUCTION.. 1

1.1 Introduction.. 1

1.2 Objective of the Study 3

1.3 Significance of the Study... 3

1.4 Motivation of the study.. 3

1.5 Organization of the study... 5

CHAPTER 2 6

REVIEW OF LITERATURE... 6

2.1 Introduction... 6

2.1.1 Literature review for determinants of child schooling 6

2.1.2 Literature Review for Comparison the Classification Techniques............ 11

2.2 Summary.. 14

CHAPTER 3 15

METHODOLGY AND DATA VARIABLES 15

3.1 Introduction.. 15

3.2 Multinomial Logistic Regression (MLR): 15

3.2.1 Odds, Odds Ratio, and the Logit Transformation 16

3.2.3 Multinomial logistic model 17

3.2 Classification Tree 18

3.2.1 Splitting Criteria and Impurity Measures 20

3.2.2 Ensemble Method.. 21

3.3 Linear Discriminant Analysis 22

3.3.1 LDA Model....23

3.3 Evaluation of Classification Algorithms....24

3.4 Data And Variable....28

3.4.1 Variable Description....29

CHAPTER 4....31

RESULT AND DISCUSSION....31

4.1 Introduction....31

4.2 Multinomial Logistic Regression:....31

4.2.1 Confusion Matrix of MLR (4-9) Through Training Dataset....39

4.3 Classification Tree....39

4.3.1 Classification Tree for Child Age Group Between 4-9....41

4.3.2 Confusion Matrix of CT Through Training Dataset....43

4.4 Random Forest....44

4.4.2 Prediction from Random Forest by Using Testing Dataset....46

4.5 Linear Discriminate Analysis for Child Age Between 4-9....47

4.5.1 Result of Linear Discriminate Analysis....48

4.5.2 Plot of LDA Model....49

4.5.3 Confusion Matrix by Using the Training Dataset....50

4.6 Evaluating the Above Techniques on the Basis of Confusion Matrix and ROC Analysis....50

4.6.1 Accuracy Based on Confusion Matrix for All Techniques....51

4.6.2 Accuracy Measurement Based on ROC Analysis....51

4.6 Multinomial Logistic Regression for Child Activities Between age 10-14....53

4.6.1 Confusion Matrix of MLR(10-14) through Training Dataset....65

4.7 Classification Tree for child Age Between 10-14....65

4.7.1 Confusion Matrix of CT (10-14) by Using Training Dataset....70

4.7.2 Prediction by Using Testing Dataset....71

4.8 Random Forest of Child Age Between 10-14 71

4.8.1 Results of Random Forest 71

4.8.2 Confusion Matrix of RF(10-14) by Using the Testing Data 74

4.9 Linear Discriminate Analysis 75

4.10.Confusion Matrix of LDA(10-14) by Using the Training Dataset 77

4.10 Evaluating the Above Techniques on The Model of Child Activities Between the Age 10-14 78

4.10.1 Accuracy Based on Confusion Matrix for all Techniques 78

4.10.2 Accuracy Measurement Based on ROC Analysis 79

CHAPTER 5 82

CONCLUSION AND SUMMARY 82

5.1. Policy Recommendation 83

REFERENCES 85

LIST OF TABLE

Table 4. 1: Results of only schooling vs no schooling no work 36

Table 4. 2: Results of only work vs no schooling no work.. 38

Table 4. 3: Confusion matrix of MLR (4-9) 39

Table 4. 4: Complexity table for child activity between age 4-9................................. 40

Table 4. 5: Confusion matrix of CT.. 44

Table 4. 6: Results of random forest... 45

Table 4. 7: Confusion matrix of RF by using test data 47

Table 4. 8: Confusion matrix of LDA by using training dataset 50

Table 4. 9: Accuracy rates of confusion matrix for all techniques 51

Table 4. 10: Results of only schooling vs no schooling no work 57

Table 4. 11: Results of only work vs no schooling no work.. 61

Table 4. 12: Results of work with schooling vs no schooling no work 63

Table 4. 13: Confusion matrix of MLR for training dataset... 65

Table 4. 14: Summary of classification tree 66

Table 4. 15: Information of internal nodes 66

Table 4. 16: Confusion matrix of classification tree for training dataset..................... 71

Table 4. 17: Summary of random forest model 73

Table 4. 18; Confusion matrix of RF(10-14)... 75

Table 4. 19: Confusion matrix of LDA for training dataset 78

Table 4. 20: Accuracy rates of confusion matrix for all techniques 78

LIST OF FIGURE

Figure 3. 1: Elements of a Decision Tree 18

Figure 3. 2: Confusion Matrix 25

Figure 3. 3 : ROC Curve 26

Figure 4. 1: Decision tree; child activity for age 4-9 43

Figure 4. 2: Error of random forest for different classes using training data............... 44

Figure 4. 3: Plot of important variable... 46

Figure 4. 4: Plot of LDA 49

Figure 4. 5: Multi-ROC analysis of MLR model.. 51

Figure 4. 6: Multi- Roc curve of classification tree 52

Figure 4. 7: Multi- Roc curve of LDA 53

Figure 4. 8: Classification tree for 10-14 years child activities 70

Figure 4. 9: Error of random forest for different classes using training data............... 72

Figure 4. 10: Plot of important variable.. 74

Figure 4. 11: Plot of LDA function... 77

Figure 4. 12: Multi-ROC analysis for MLR model 79

Figure 4. 13: Multi-ROC analysis for CT model.. 80

Figure 4. 14: Multi-ROC analysis for LDA (10-14)... 81

ABSTRACT

This study investigates the factors which are more affective in the decision of child activities through classification techniques. Further, it compares the classical approach and machine learning techniques of classification on the basis of overall accuracy of confusion matrix and area under the curve (AUC). The data is taken from Pakistan social and living standards measurement (PSLM) survey for year 2014-2015 and is based on urban and rural areas of four provinces of Pakistan. Two separate models are made based on age groups 4-9 and 10-14. Results showed that accuracy from confusion matrix and area under the curve ROC analysis of classification tree model is greater than the MLR and LDA for the age group 4-9. While accuracy from confusion matrix of classification tree is greater than MLR and LDA for age group 10-14. However, accuracy checked in the context of area under the ROC analysis showed no significant difference between the accuracy of three model. Our finding show that classification tree is best technique among others as it also identifies the most significant variables. Such as ,child gender, kaccha house, fuel for cooking mother education ,mother employment,region ,child'sage, infants, toilet facility, aggland, cattle, 16-64 female, source of drinking water and father employment. Therefore, it is recommended to reducing the gender discrimination towards child activities. Gender disparity should be minimized through public awareness about girls' education. Woman education in both model have significant effect on the decision of child activities. We have to focus on girl's education because in future girls can play important role as woman. It has an increasing effect on human capital through the education.

CHAPTER 1

INTRODUCTION

1.1 Introduction

"Education is the movement from darkness to light" (Allan Bloom)

Economic progress of developing country as well as developed country depends on two type of accumulating the capital that is physical and human capital .the human capital accumulation has its own importance's . human capital investment has positive effect on economic growth. Schultz and Becker in 1960's has verify that the step of economic progress of developed countries is impossible without the investment in human capital. The expansion of human capital or resources is incredible without stress on education.

Human capital investment in education is the building of any economy. We all know that educational investment in children improve their future earning capacity and career opportunities as well .other benefits are also attach with the educated child's like they have ability to get new knowledge , improve productive skills , improve health status and help to reducing poverty level etc. All these benefits are not limited only individual level they can transfer to the family and then economy level.

In many developing countries like ,Pakistan has a severe problem of human capital .majority population of our children are not engage only schooling activity .they may be engage in work activity(child labor). In Pakistan, children are making extraordinary economic contribution to their family.so, it is also claimed that there is compromise between work and schooling because pushing child in economic or a productive activities might increase current income and improve living standard but will extremely challenge her human capital (education) growth.

However, schooling and work (child labor) are not certainly perfect inversely associated in time distribution. Many children are engage in both activities (schooling with work) i.e. child attend schooling in morning and after schooling they do part-time job. There is clear evidence that the increasing trend in schooling is not reduced labor force participation .there is also possibility that children cannot engage in any schooling or work activity ,it may be idle.

Pakistan facing many problem in the accumulation of human capital through education .policy on education has been clearly deficient due to many reason. Pakistan can attaining the goal of free and necessary education for every child so they should make more policies on education .for the purpose of better policy we should know about child activities and also know about which factor is effecting in child schooling and alternative activities.

Therefore, we predict the child's activities through econometric model of classification with the help of both classical (traditional) and machine learning techniques of classification . econometries can usually use classical approach (logistic regression) for classify the data.in this techniques statistician emphasis on conditional distribution of y given some other independent variables x.

Now a days we have available lot of data for analysis and manipulating so large dataset may have create complex relationship. For complex relationship in large dataset we need more controlling tools for manipulating the data. In the large dataset modeling and estimating the complex relationship we required the machine learning techniques. machine learning techniques is mostly concerned with prediction. These techniques are semi-automated extraction of information from data. Semi-automated mean that it can involve many keen decision by a human. Machine learning can also suggest tools for detecting and summarizing the nonlinear relations in data and also discovery some function which give better prediction for y as a function of x. these nonlinear techniques are classification and regression trees(CART) , random forest ,support vector machine ,penalized regression etc. Here we used CART and random forest techniques .

Economist and many researcher would usually consider the logistic regression for classify the data but if we have lot of data for manipulating than better techniques are available in machine learning

1.2 Objective of the Study

The objectives of the study are to;

- Examine the importance of socio economic factors in determine child schooling and other activities by using the Multinomial logistic regression, Classification and regression tree and linear discrimination analysis.
- Compare classifier techniques with each other and examine which classification techniques can work better on the bases of confusion matrix and ROC analysis(area under the curve)
- Identify the main factors of child schooling and other activities with the help of decision trees

1.3 Significance of the Study

Illiteracy is common issue among developing countries; infact in most of these countries child education is compromised with the child labor for increasing household financial resources. Therefore, it is necessary to identify key demographic and household determinants in defining child schooling and alternative activities of child.

1.4 Motivation of the study

Education is a basic component in the development of a human character, knowledge and future. It is a major human capital which elevates poverty and removes income inequality from the economy by improving the quality of life and increasing the total factor productivity. A study on agricultural productivity in Pakistan shows that four years of schooling on average increases the output of farmers by 8 percent. A 10 percent increase in male literacy in Pakistan will lead to increase the agricultural productivity by 2.7% [UNICEF 1997].

Unfortunately in Pakistan, like many other developing countries children either do not have access to education or are enrolled in schools of questionable quality (Khandker ,*et al* .1994). Parents only provide education to those children which they think are bright. Moreover, poor children have higher IQ levels but they are unable to attend school as they have to earn for their families. Government should need to devote more resources to education sector in order to improve access to primary and secondary education. This is evident from the UN's Millennium Development Goals, the second and third of which, respectively, aim to "ensure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling". This can only be done effectively by having an insight into the factors that determine the schooling outcomes of children in a household as well as those that impede children's participation in schools. The general phenomenon in developing countries is that the decision to go to school is intimately related to the decision to work. There are number of factors such as parents' education, their employment and health as well as the child's age and the number of siblings and their age composition and the relative level of household poverty are important demand side factors affecting the decision to go to school or drop out. Studies on schooling decisions have investigated a number of determinants for low levels of participation in primary schools and high rates of dropout.

1.5 Organization of the study

This study is organized into five chapter. Chapter one contains general introduction, objective of the study and significant of study. In the second chapter provides the literature review. three chapter concentrated on the econometric methodology, description of variable and sources of variable. In chapter four we provide the results discussion and in the chapter five we provide the conclusion and summary of the study.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Introduction

As mentioned in the first chapter, the child education is an important factor. A number of studies had explained the link between child schooling and child labor and had indicated the socioeconomic factor which affects the child schooling on a broader scale. This chapter reviews these studies for explaining the issue with technical research papers.

2.1.1 Literature review for determinants of child schooling

Sather &Lloyd, (1994) have looked on the major issues in Pakistan that is who get primary education in the framework of "inequality among and within families". They investigated the determinates of parental decision for child schooling, using PIHS survey of 1991.their finding showed that inequalities are present across the household that are main description for disparities between children in completion of primary schooling and inequities are present within household show disparities between gender. They also find that parental education especially mother education can play important role in the decision for children attend school and complete primary education

Jensen &Nielsen, (1997) explored the factors affecting the choice of child schooling and child labor .they collected the survey data from Zambia and estimated the determinate of child schooling decision through logit model. They found that poverty forces the households to retrain their children from school.

Burki & Fasih, (1998) explore the determine of child labor in Punjab ,Pakistan and also worked on non-leisure time allocation for children .they investigated this issues by using child labor survey 1996 .through multinomial logit regression and they create dummy of child ages in the child activities model .

Duraisamy, (2000) examined the determines of child schooling and work contribution of girls and boys. They investigate through probit and multinomial logit regression by using the survey data of NCAER 1994.results showed that parents education especially mother education and household income had significant impact on children attending the schooling and discouraged the child participation in work

Cockburn, (2001) studied child labor versus education in the context of "poverty constraints or income opportunities" .They used multinomial regression on data from Ethiopian agriculture household modeled determinants the household income and demand for child labor . Results showed that both factor(poverty constraints or income opportunities) is important in the decision of child schooling or not

Blunch, *et al*. (2002) have looked on the recent empirical studies of "participation of children schooling and labor activities" in the selected developing countries. They explore that poverty ,parental educational and employment status ,age and gender are had significant on the school attendance and child labor .

Heltberg & Johannesen, (2002) worked on the paper about "how paternal education can effect on human capital" for this purpose they used four indicator of human capital .they analyzed this paper by using survey data from household of Mozambique through sequential model .they encounter that parental education especially mother education can play significant role on the education ,health and fertility.

Khan, (2003) explored the different activities of children's. For this purpose they used cluster sampling techniques to collect the data from two district of Pakistan that is Pakpattan and Faisalabad. They estimated the model through sequential probit model. Finding showed that parents education had positive impact on the decision of child schooling especially mother education .Also, they found that birth order had significant impact on child schooling and labor ,children participation reducing when increasing the child age.

Ali & Khan, (2004) examined the supply side determines in urban area and also investigated how child education, household income ,parental education ,unemployment level of parent's and demographic variables child labor .they investigated through sequential probit model by using the field survey consist 2000 sample of urban household which are collected from district Pakpattan.

Khanam, (2004) by using the survey data collecting from Bangladesh inspected the decisions of households are involving in child activities (schooling and work).They issue estimated through multinomial logist regression and also checked the impact of work on the school attendances and school achievements .Results showed that parental education had significant impact on increase the probability of child schooling and if father is on daily wages then chances to increase that child are involving work with schooling activity. They also found that girls school attendances and grades achievements are more effected than boys due to the more involving in work with schooling activity than boys.

Parikh & Sadoulet, (2005)investigated the effect of parents employment status on the child labor and schooling in brazil .Finding show that self-employed parents are more involved in the work activity. Furthermore results showed that children are more likely to be involved in work activity where average of adult child employment is high.

Nkamleu & Kielland, (2006) discovered the famers decision on the child labor and schooling in the coca sector by using the survey data collected from coca household. They estimated the model through multinomial logistic regression have four categories and found that child farming in coca sector and nonenrolment both are significant.

Hussain, *et al* .(2008) investigated the socio economic factors which affects the parents decision to public and private child schooling at primary level in Punjab, Pakistan used HICS data. They explored the determinants of selecting child schooling by using binary logit model .Finding show that the reason of higher enrollment of child in private schooling is the quality of education the trend of private schooling is more in urban area. They also found that family size ,child age, schooling cost, distance from school had negative effect on private schooling

F.N. & Yinusa, *(*2009*)* by using labor force survey of 2005-2006 explored the "determines of child schooling and child labor in Botswana*"* through multinomial logistic regression. Results indicated that female head household ,head of household employment status and child age had negatively affected on the probability of child work with schooling.

Khan, *et al*. (2010) did a the comparative analysis of child labor determines in rural and urban area of Pakistan .For this purpose they collect the sample from two district of Pakistan and investigated household decision through sequential probit model. They explore that parental employment status had positive impact on child schooling in urban area but negative in rural area while especially in the scenario of mother employment had positive impact on child labor in rural area but negative in urban area and poverty. As per the results gender discrimination for child schooling was higher in rural area.

Oni, *et al*. (2010) studied that determinant of child labor and schooling in rural northeastern Nigeria. The empirical results indicated that 54% of children feed their families along with schooling and while 5.9% were not working or schooled. The study explored yet another important aspect which largerly affects the child schooling , the health condition of the head of the household.

Yamada, (2011) checked the long term impact of family background on the educational achievements ,family creation , labor market consequences and spousal features for Japanese woman ,for investigated this issue by using 'Japanese Panel Survey of Consumers "(JPSC) from 1993 to 2004.they found that those who have less number of sibling are increase their education and those belong to rich family then trend to child get private education

Olanrewaju &Olaniyan, (2011) estimated the household and individual determinants of child schooling by using survey data taken from the 1999 Multiple Indicator Cluster Survey (MICS) .They explored through the probit model and found that socioeconomic backgrounds of children had important determinants. Educated parents had promoting more schooling and also pinpointed that the gender gap in a rural area are more than the urban area.

Lodhi, *et al*. (2011), reported the effect of households, individual and community level factors on the possibility of children involved in various activities. They investigated the issue by using the survey data constructed on the interviews of 40 different villages of four Pakistani provinces through Multinomial Probit model. The result showed that the parental awareness had played a major role in the possibility of involvement in secular school attendance, religious education, and child labor. They also found that in rural area female child is more involved in child labor activity as compared to the involvement in secular schooling and also found that parents focus more on male child schooling than female child.

Ahmad & Hussain, (2012) studied those two main issues in labor market activities, the first issue related to the youth behavior towards work and education. Second, issue related to the supply-side determinants of youth activities in Pakistan. There analyses made use of microdata from Labor Force Survey (2006-07) and used Multinomial logit model. The result showed that educated parents prefered more schooling while the parents working in informal sector like agriculture sector preferred to work instead of schooling. Due to the large size of household, the share of a child in economic participation is increased.

Qureshi, *et al*. (2014), presented a paper about Child Work and Schooling in Pakistan. They investigated the determinants of non-income factors like household socioeconomic, parental background, demographic and examined the child schooling and child labor nexus by using the Panel Household Survey 2010 dataset through the probit model

Iddrisu , *et al*. (2017) studied the determinants of child school enrollment in Ghana. They investigated this issue by analyzing the survey data Collected from GLSS 6 using sequential logit model. The result showed that family resources like parental education household income and the gender pf HH head played important role in the child schooling decision. They found that educated mother and father enrolled their children in school without any gender biased discrimination.

2.1.2 Literature Review for Comparison the Classification Techniques

Press & Wilson, (1978) debated on choosing between two method of classification ,linear discriminate analysis and logistic regression .they found that if LDA assumption are fulfil than its better than logistic regression however if one variable is not multivariate normal than prefer the logistic technique .

White, (1987) classify the cow having mastitis or not for by using two classification method ,logistic regression and linear discriminate analysis .they compared the two techniques on the same dataset .Results indicated that coefficient of logistic regression were better than LDA because of assumption of LDA method were not fulfilled.

Worth & Cronin, (2003) worked on human heath effected by different type of chemical .they classified the different type of chemical through three techniques of classification logistic regression ,classification tree and linear discriminate analysis. they concluded that classification tree is appropriate method.

Blas, *et al.* (2004) on the bases of many measures of predictive accuracy compared the logistic regression and linear discriminate analysis via simulation. Main purpose was to choose between two techniques and set some standard for choice of techniques .results indicated that when normality assumption are not violated ,both techniques gave all most same results.

Song, (2008) compared the classification methods on the credit card approval data by using six classification technique. They also identify that which variables are the important factor to decided approval of credit card. Classification and regression tree and logistic regression performed well in the credit card data.

Panagiotakos , *et al* .(2009) estimated the logistic regression and linear discriminate analysis for the Evaluating Factors Associated with Asthma Prevalence among 10- to 12-Years-Old Children. They also evaluate the convergence of both methods. Results showed that both method predicted the symptoms of asthma among Greek children had same level of convergence and similar results

Tabriz, *et al.* (2010) investigated role of human factor in incidence and severity of road crashes in Iran. They analyzed traffic data by the data mining techniques such as logistic regression and classification and regression tree. The result indicated that the human factors such as Driving license and Safety belt are important role in the severity of accidents in Iran.

Young hu, (2011) classified the four types of financial fraud by using the four techniques of data mining all are detected the categories of fraudulent data. research are also address the gap between financial fraud data and need the industry to inspire further research on ignored areas.

Mahjub, *et al.* (2013) compared the six classification techniques on the real data of diabetes .they classify that person is diabetic or not .They compare the techniques on the bases of sensitivity ,specificity ,total accuracy, and area under the curve getting from ROC analysis. They found that in terms of total accuracy and area under the curve, support vector machine technique is superior than the other on diabetes data.

Kanwal, (2016) worked on the comparison between three classification techniques that is logistic regression ,linear discriminate analysis and classification techniques on the real world factor of expecting woman heath. They used survey data collecting from PLMS Islamabad and checked the techniques performances on the bases of ROC analysis (area under the curve).Results indicated that classification tree technique was better than other two techniques.

Khan,*et al.*(2014) estimated the classification tree ,logistic regression and linear discriminate analysis for the classifying the data of water quality. The data on water quality were obtained from the Pakistan Council of Research in Water Resources (PCRWR) for two cities of Pakistan. The result indicated that logistic regression can performed well as compare to the other two techniques .The linear discriminate analysis and classification tree performed equally but interpretation of classification tree were comparatively easy

Akingbade, *et al*. (2015) checked the performances of logistic regression and discriminate analysis on the data of "delivery of an expectant mother". they identified in linear discriminate analysis , mother weight and age are important variable in the mode of delivery. Results are also indicated that both methods in term classification rate were same but LDA assumption were not followed so they prefered the coefficient of logistic regression.

Ali, (2015) compared the three classification techniques on the German credit data. Logistic regression, classification and regression tree and random forest were used to classifying the loan application in to the good loan and bad loan. They also investigated that the performance of each three classification techniques on the basis of accuracy, precision, negative predictive value, recall and specificity. Classification and regression tree performed best in the accuracy, specificity measures and precision, logistic regression has best performance at the low probability of default thresholds while random forest performed best for negative predictive value and recall measures.

Mezerji,*et al*.(2015) estimated the logistic regression and linear discriminate analysis for the prediction of depression in cancer patients. Their analysis are based on the cross-sectional study selected 243 cancer patients .They compared LR and LDA models using the classification indices. The results indicated that LR perform better in some cases and LDA in other on the bases of classification error (CE) .classification error index is not suitable for the classification other indices B and Q better performed and more efficient criteria for the comparison.

2.2 Summary

We conclude that different researcher estimated that the determinants of child activities by using the classical approach like logistic regression but no one can be used machine learning techniques such as, LDA(linear discriminate analysis) and CT(classification tree) on the theory of child activities. Pervious literature is evidence that when we classify the dependent variable use different classification techniques such as LDA and CT, so we used machine learning techniques and compare with the classical approach.

CHAPTER 3

METHODOLGY AND DATA VARIABLES

3.1 Introduction

This chapter discusses the classification techniques of estimation which will be used to examine the importance of socio economic factors in affecting child schooling. The section 3.2 provides the detail on Multinomial logistic regression, Classification tree, and Linear discriminate analysis techniques along with their assumptions. Section 3.3 discusses the methodology of tools used for the evolution of classical and machine learning techniques.

3.2 Multinomial Logistic Regression (MLR):

Logistic regression (LR) comes under the classification techniques introduced by cox (1958). This technique is commonly used to measure the relationship between categorical dependent variable and more than one independent variables. Logistic regression may consider binomial or multinomial dependent variable. In Binomial logistic regression a the dependent variable have only two outcomes such as yes /not, true / false, male / female, win / alive etc. Whereas, in multinomial logistic regression the dependent variables have more than two possible outcomes, for example, as in this study we are taking child activity as a dependent variable having four possible outcomes, i.e. no schooling no work, only schooling, only work, and work with schooling. Multinomial logistic regression is helpful to predict probabilities of more than two outcome of the categorical variable.[1] Logistic regression is nonlinear technique.

3.2.1 Odds, Odds Ratio, and the Logit Transformation

Consider the regression model

$Y= \beta_{\circ}+ \beta_1 X_1 \ldots\ldots + \beta_k X_k+ \varepsilon$

In matrix form,

$$Y= X\beta+\varepsilon \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (3.1)$$

where Y is dependent variable categorical in nature and X is vectors of independent variables consist of may be categorical and continuous variable, let Y has two categories 0 and 1. In order to predict the probability to classify that whether the depended variable will be 0 or 1,the conditional probability of Y=1 given the independent variable x, P(Y=1/X=x) equal to π(x) is define as where π(x) is the conditional probability of success which lie between 0 and 1 and P(Y=0) can be found as P(Y=0/X=x) = 1-P(Y=1/X=x) =1- π(x) Once we find the probability ,the odd ratio is define as Odd ratio (g(x)) is equal to the $\frac{p(sucess)}{1-p(sucess)} = \frac{P(Y=1)}{1-P(Y=1)} = \frac{\pi(x)}{1-\pi(x)}$ applying the natural log on odd ratio it's will produce a logit of Y i.e,

[1] The situation where we predict more than two outcomes also called polychromous, multiclass, polychotomous logistic regression, maximum entropy classifier and conditional maximum entropy.

Logit (y) = ln{$\frac{\pi(x)}{1-\pi(x)}$} , the relationship between odd ratio and logit (Y) Odd ratio = $e^{logit(Y)}$ and equal to the $e^{\ln[odd(Y=1)]}$ when we find out the probability of Y=1 , P(Y=1) =$\frac{\pi(x)}{1+\pi(x)}$ and π(x) equal to the $\frac{e^{\beta_\circ+\beta 1x1}}{1+e^{\beta_\circ+\beta 1x1}}$, so $e^{\ln[odd(Y=1)]}$ $= e^{\beta_\circ+\beta 1x1}$

3.2.3 Multinomial logistic model

In the multinomial logistic regression, outcome variable(Y) has more than two categories. .Let's assume three categories of Y and are coded as 0, 1, 2, .In the three outcome category model we have two logit function Therefore , in the multinomial depend variable first we have to decided that which category is a reference category or a baseline outcome. Suppose that Y = 0 is the reference category and to form logits comparing with Y = 1 and Y = 2. Two logit functions as

$$g_1\text{ (x) = Ln}[\frac{P(Y=1/x}{P(Y=0\ /x}]$$

$$= \beta_{1_\circ}+\beta_{11}x_1+\beta_{12}x_2\ldots\ldots\ldots+\beta_{1k}x_k$$

$$= X^{'}\beta_1 \qquad (3.2)$$

$$g_2\text{(x)= Ln}[\frac{P(Y=2\ /x}{P(Y=0\ /x}]$$

$$= \beta_{2_\circ}+\beta_{21}x_1+\beta_{22}x_2\ldots\ldots\ldots+\beta_{2k}x_k$$

$$= X'\beta_2 \qquad (3.3)$$

Conditional probability of each outcome category gives the covariate vector as

$$\text{P (y=0/x) =}\frac{1}{1+e^{g1(x)}+e^{g2(x)}}$$

$$\text{P(y=1/x) = }\frac{e^{g1(x)}}{1+e^{g1(x)}+e^{g2(x)}}$$

$$\text{P(y=2/x) =}\frac{e^{g2(x)}}{1+e^{g1(x)}+e^{g2(x)}}$$

Interpretation in the linear regression model is easy as compare to the logistic regression model. In the linear regression model, we interpret the slop coefficient in a way that the change in dependent variable when independent variable changes by one unit and all other independent variables are held constant. In the logistic regression model, w co-efficient cannot be interpret as in the liner regression model because in this model the link function is involved that is the logit transformation P(Y=1/X=x) $= \frac{\pi(x)}{1+\pi(x)}$ therefore, we interpret slope coefficient in a way that the change in logit corresponding to change in one unit in the independent variable all other variable are held constant.

3.2 Classification Tree

Classification and regression tree (CART) is introduced by Leo Breiman (1984).the CART algorithms is referred to as a decision trees. Decision tree builds classification models in the form of a tree structure; it breaks down a dataset into a smaller and smaller subset. Elements of a Decision Tree are, root node is the parental node for root node there is no incoming edge but it has outgoing edges. At root node we have all the predictor's space X. Internal node or Non-Leaf node that point of tree where the predictor space splits is referred as internal node and Leaf or Terminal node represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. We make prediction at leaf node and average all the training data points which belong to that leaf. All process is show in the Figure 1

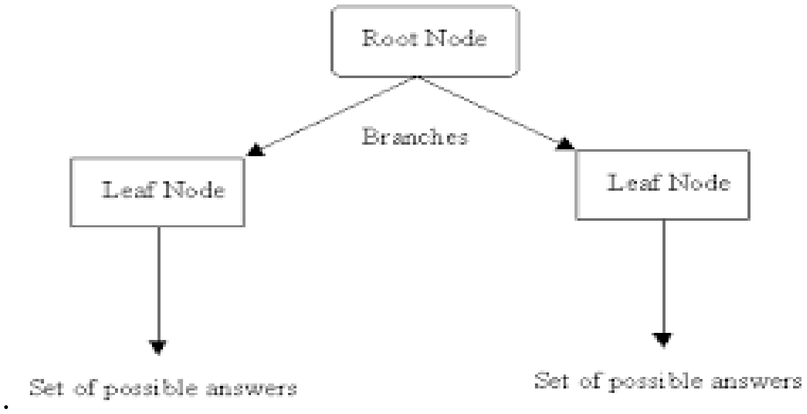


Figure 3. 1: Elements of a Decision Tree

There are two types of decision tree one is classification tree and another regression tree. Classification tree is used for qualitative variable whereas, regression tree are used for quantitative variable. CART tree is basically a binary decision tree because it is constructed by splitting nodes into two child nodes and this process is repeated until the tree growing process is stop. This process is start with the root nodes. The tree growing process is to pick a split between all other possible splits at each node so that the resulting child nodes are purest. In CART algorithm, only univariate split are considered each split depends on the value of only one X (predictor) variable and each independent variable have many possible splits. If predictor or independent variable is nominal (more than two categories or n categories) there are 2^{n-1} - 1 splits and if predictor is continuous variable (having infinite number of possible values) let say continues variable with g unlike values then there are g -1 possible splits. Tree growing process start from the top node called root node by frequently using the following steps on every node.

- First sort the values of all predictor variable both nominal and continues in ascending order (smallest to the largest).then examine the each sorted predictor go through each value from the root node to examine each candidate split point (w), if x ≤ w in this case child node goes to the left side, if not then goes to the right side.

  For each sorted nominal predictor, go through each possible subset of the categories (if subset is R) if x belong to R in this case child node goes to the left side, if not then goes to the right side this process is help out to find best splits.
- After finding the best splits then next find out the nodes best split, for best node split, choose the one that maximizes the splitting criterion.

- Best nodes split process is continue until when the stopping rules are not satisfied.

Tree growing process are continues to grow or not depend on the stopping rule, following stopping rules are used

- If all cases, in a node have same values of dependent variable then the node cannot split further, nodes become “purest”.
- If all cases, in a node have same values of each independent variables (predictors) then node cannot split.
- If the tree is reaches to the user specified maximum tree depth limit value then tree growing process will stop.
- If the size of the node less than user specified minimum node size then the node cannot split.
- If in the cases of splitting nodes resulting in child nodes values and values of the child node is less than user specified minimum child node size, then node cannot split further.

3.2.1 Splitting Criteria and Impurity Measures

At node *t*, the best split *s* is chosen to maximize a splitting criterion $\Delta i\ (s,t)$. When the impurity measure for a node can be defined, the splitting criterion corresponds to a decrease in impurity. Y is categorical variable there are two splitting criteria are available, such as Gini index, and cross entropy if Y is qualitative response then splitting criteria is Residual sum of square (RSS). (RSS) cannot use for splitting notes when we use qualitative response variable so we use classification error rate (CER) as an alternative to RSS. Classification error rate is splitting note down at each internal note.it is the fraction of the training observation in that region that do not belong to the most common class. Gini index and cross entropy is two other methods for measure the error rate.

Gini index is

Gini index is generalization of binomial variances.it is define as

$$G= \Sigma_{k=1}^{k} \hat{p}mk(1 - \hat{p}mk)$$

Where $p^{\wedge}{}_{mk}$

represents the proportion of training observation in m region that are form the k class. The Gini index take on very small value if $p^{\wedge}{}_{mk}$ are close to zero and otherwise one

Cross Entropy

$$D \ = -\Sigma_{\boldsymbol{k=1}}^{\boldsymbol{k}} \hat{p}mk \log(\hat{p}mk)$$

As $0 \leq \hat{p}mk \leq 1$, it follows that $0 \leq \hat{p}mk \log(\hat{p}mk)$.

3.2.2 Ensemble Method

The CART algorithms provide a foundation for important algorithms like bagged decision tree, random forest and boosted decision trees. These techniques are more powerful for getting better predictions in decision tree. We used most common method (random forest) for ensemble method.

Random Forest is an addition of bagging technique. this technique has one additional to bagging technique step, the addition step takes the random subset of data along random selection of features not using rather than using all features to grow trees .Many random trees are called Random Forest.

Random forest takes following step

1. There are many observations and many features in training dataset. First, take the sample randomly with replacement from training dataset.

2. Selected subset of many feature are randomly and any feature gives the best split then used the best split as the node iteration.

3. Tree is grown-up.

4. All these above steps are repeated and prediction is set on the base of combination of predictions from n number of trees.

3.3 Linear Discriminant Analysis

Discriminant analysis was introduced by the fisher in 1936. Discriminant analysis is used in statistics, pattern recognition and machine learning. Discriminant analysis is basically statistical technique to classification between two or more groups with the various set of explanatory variables. Discriminant analysis is used where the cluster are known as priori. .The objective of this technique is to classify the various observation into the know groups. Discriminant analysis and Multivariate Analysis of variances (MANOVA) both are the same mathematically they are difference to each other in term of dependent and independent variable .in discriminant analysis the independent variable perform as a predictors while dependent variable determine the group membership where in the MANOVA the dependent variable perform as a predictor while independent variable determine the group membership The statistical tools of MANOVA and discriminant analysis both have a important base in the superior linear model. Both methods are observing at multivariate variances among groups.

LDA assumptions are follow

- It is assumed that the distribution is Normal (Gaussian) for all variable .you can examine the variable are normal distributed or not through the histogram of frequency distribution or we can also perform the Shapiro-Wilk test for testing multivariate normality.

- It is assumed that sample size of the smallest group required to exceed number of predictor variables and unequal sample size are acceptable
- It is assumed that the population variance and covariance's for all independent variables are equal across the dependent variable ,also known as the homogeneity of variance – covariance
- It is assumed that there must be low multicolinearity among independent variable because if one of the independent variable is highly correlated then the discriminant function coefficients will not reliably measure the relative importance of the predictor variables.

3.3.2 LDA Model

In the case of many populations, let say k population then we can use the classification technique proposed by the Fisher. LDA are basically used to separate the input data by dimension reduction and to develop the input data on a lower dimensional space in which the data of different classes are well separated as much possible. LDA is indirect approach for estimating the predicted probabilities of response variable. It builds a predictive model for group membership .The model is consisting of a discriminant function base on linear combinations of predictor variables. Predictors deliver a best discrimination between two or more groups. LDA used both conditional and marginal probabilities based on Bayes Theorem

$$\Pr(Y=K/X=x) = \Pr(\frac{X=x|Y=k).\Pr(Y=k)}{\Pr(X=x)})$$

Writes this somewhat differently as

$$\Pr(Y=K/X=x) = \frac{\pi_k f_k}{\sum \pi_l f_l}$$

Where f_k equal to the $Pr(X = x|Y = k)$, f_k is the normal density for X in class k, we will use normal densities for these, separately in each class and π_k equal to the $\Pr(Y = k)$ is equal to the marginal or prior probability for class k. The Gaussian density has in the form of

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \quad e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

Here μ_k is the mean and σ_k^2 is variance and will assume that variance are constant

Equation plugging into bayes formula we get a rather complex expression for

$$P_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}}{\sum_{l=1}^{K} \pi l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu l)^2}{2\sigma^2}}}$$

To classify at the value X = x, we need to see which of the pk(x) is largest. Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest discriminant score

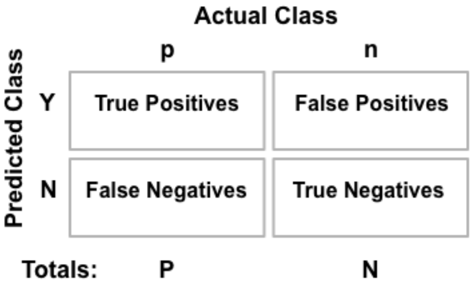
$$\delta_k(x) = x.\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Note that δ k(x) is a linear function of x

3.3 Evaluation of Classification Algorithms

Evaluation the classification techniques are the one of the important topic in any method of data mining .The most commonly tools used in evaluating the results of classification algorithms applied are: confusion matrix, and receiver operating curves (Oprea,2014). Confusion matrix is a table in which show the number of incorrect and correct predictions made by the model compared with the actual classifications in the test data.

For the simplification we start classification problems with only two classes. This type of matrix are 2x2 confusion matrix and also called the contingency table with two dimensions “actual” and “predicted” ,and matching sets of classes in both dimensions .to separate the classes (actual and predicted) we use labels {Y,N } and p and n stand for positive and negative .see in the below figure.



		Actual Class	
Predicted Class		p	n
	Y	True Positives	False Positives
	N	False Negatives	True Negatives
Totals:		P	N

FIGURE 3. 2: CONFUSION MATRIX

In this Figure 3.2, there are four possible outcomes true positives, false positives, false negatives, and true negative .if prediction is Y and actual class is also Y it is counted as true positive, if prediction is Y and actual class is N it is counted as false positive, if prediction is N and actual class is Y it is counted as false negatives and if in the case prediction is N and actual class is also N it is counted as true negative .the outcome represented in the diagonal is the correct decision and off diagonal is represented the error or confusion between the classes. True positive rate (TP) is also called the hit rate or recall rate.

$$\text{TP} \quad \text{approaches to} \quad \frac{Positive\ correctly\ classified}{Total\ Positives}$$

False positive rate is also called false alarm rate

$$\text{FP approaches to} \ \frac{negatively\ incorrectly\ classified}{Total\ negative}$$

Other terms related with ROC curves the terms are sensitive and specific. Sensitive are equal to the true positive rate (recall rate) and specific equal to the

$$\frac{true\ negative}{false\ positive+true\ negative} = \text{1- false positive rate}$$

Same procedure can be apply on 3x3 and 4x4 confusion matrix

ROC Curve

ROC stands for receiver operating characteristic curve.it is graphically technique use for the selecting classifier based on their performance. ROC graph are usually used in medical field for decision making purpose, in recent year ROC curve are used in machine learning and data mining. ROC graph are 2 dimensional graphs in which sensitivity (true positive rate) is on Y axis and 1-specificity (false positive rate) is on X axis .in this graph the tradeoff between sensitivity (benefits) and 1- specificity (cost) .in ROC space figure with four classifier label as A,B,C,C. show in the figure 2

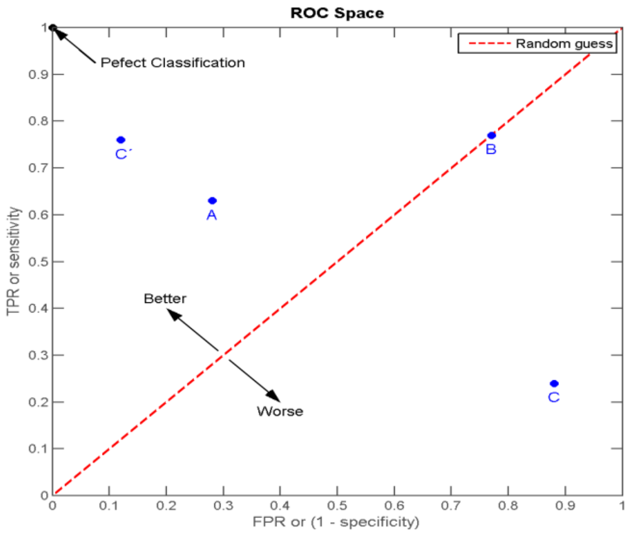


FIGURE 3. 3 : ROC CURVE

Every classifier is producing a single pair in the roc space with the corresponding of FP rate and TP rate. Some point are important in the roc space such as (0,0) ,(0,1) ,(1,1).the (0,0) point is located in the lower left area this point represented that no false positive error but cannot gain true positive .the (1,1) point is also located in the right sight this point is worse side and (0,1) point is located at the upper left corner this point represented the perfect classification (0,1) point is the better than other points . Roc curve is the two dimensional representation of classifier performance .To compare classifiers we may want to reduce ROC performance to a single scalar value representing expected performance for this purpose common method is calculate that the area under the curve (AUC), value of AUC is between 0 and 1 .in the above figure diagonal line between (0, 0) and (1, 1) produces the random guess. AUC has an important statistical property that the probability that a classifier will rank randomly chosen positive occurrence higher than negative one. Roc curve is also used for the more than two classification problem this situation is more difficult if whole space is to be managed let’s suppose n classes, in the n x n confusion matrix contain the n outcome represented in the diagonal is correct classification and n^2 – n outcomes in the off diagonal represented the incorrect classification .with 3 x3 confusion matrix

In this confusion matrix we have 3 correct values (benefits) and in the three classes the surface become (3^2- 3) 6-dimensional. For handling the 3 or more classes let say n classes one method is to make n different ROC graph ,for every class . Call this the class reference formulation. Specifically, if C is the set of all classes, ROC graph i plots the classification performance using class $c_i$as the positive class and all other classes as the negative class .The area under the curve is used to measure the discriminability of a pair of classes. AUC is usually use in two class problem (or a single scalar values) but they can be used in Multi class problem.in this case introduce the issue of combining multiple pairwise discriminability values .one way to calculating AUC for multi class by generating each class reference ROC curve in turn, measuring the area under the curve, then sum up the AUCs weighted by the reference class occurrence in the data.

3.4 Data And Variable

The current study is using Pakistan social and living standards measurement (PSLM) survey for year 2014-2015. PSLM is considering both urban and rural areas of four provinces of Pakistan. In this study, we are take child activity as a dependent variable and which are further categories in to four categories i.e., no schooling no work, only schooling, only work and work with schooling. We take the sample where child age is between 4-14. In this study we make two models for determining the child activity based on two age groups that is age group between 4-9 and age group between 10-14 as the information on work with schooling category is not available for age group 4-9. Therefore, In the first model we considered child activities for 4-9 ages as a categorical variable .In this model child activity has three categories, these categories are no schooling and work, only schooling, only work and In the second model we considered child activities for 10-14 ages as a categorical variable. In this model dependent variable has four categories there are no schooling and work, only schooling, only work and work with schooling. Explanatory variables for the determining the child activities are grouped into three categories such as demographic variable, household socioeconomic variable and household parental background variables. These variables have been selected on the basis of pervious relevant literature.

3.4.1 Variable Description

In our study we used many predictor which are selected through the pervious literature those variable are divided into three groups .these group are demographic variables , household socioeconomic variables and household parental background variables.

First we discussed about the household parental background variables group, in this group we have information about child’s father and mother education ,child’s father and mother occupation .**Child’s father education** :have two categories Yes or No if a person able to read or write we will consider the person is educated so we report the Yes if not then reported the No

Child’s mother education : Having two categories Yes or No

Child’s father occupation : having four categories ,Father employee ,Father paid employee ,father unpaid worker ,father self-employment

Child’s mother occupation : it have five categories, mother employee ,mother paid employee ,mother unpaid worker ,mother self-employment and mother unemployed.

second we disused about the demographic variable in this group we have information about child gender ,region, child age, numbers of female family member between age 16-64, numbers of male family member between age 16-64,numbers of elder, numbers of infants .

child gender: it is categorical variable and having two classes ,Male and female

Region : is having two categories Urban and Rural

Child ages: is continuous numeric variables**,** we have taken the samples of child ages between 4-14

Numbers of elder: we have taken the sample of elder is equal to and greater than 65 age

Numbers of infants: we have consider the infants whose age is less than 4

We discussed about the demographic variables in this group we have detail information about the source of drinking water, sanitation facility ,main fuel use for cooking ,kaccha house ,one room house ,personal house, numbers of cattle , agriculture land in hectares

Source of drinking water :dividing into two categories that is Piped water and other source of water(hand pump, open well, covered well, river ,stream, etc.)

Sanitation facility: it is categorical variable and dividing into two categories that is Facility not available and other toilet facility

main fuel use for cooking :its divided into two categories that is Gas and other fuel .in the other fuel categories is included fire-wood, crop residue ,coal ,kerosene oil)

kaccha house :it can divided into two categories kaacha house and no kaccha house

personal house: dividing into two categories that is Yes and No

numbers of cattle: is the numeric variable

agriculture land in hectares :is the continuous variable

CHAPTER 4

RESULT AND DISCUSSION

4.1 Introduction

In this chapter use to estimate the model of three classification techniques ; classification tree, multinomial logistic regression and linear discriminate analysis and evaluated these techniques to check which one is the best in the empirical analysis of child work and activities. For estimation we spilt the dataset into two non-overlapping groups ; training and testing set. The training sets contain 60 percent data and testing test contain 40 percent data. After estimation we compare the performance of these classification with the help of prediction accuracy and based on confusion matrix and sensitivity specificity through ROC analysis .

Following the introducing the rest of the chapter is divide into Section. Section 5.2 results and discussion of estimated parameters for children belonging to age group 4-9 years and evaluation measurements of the estimated classification techniques. Section 5.6 represent the result of estimated parameters for children belonging to age group 10-14 years and evaluation measurements .

4.2 Multinomial Logistic Regression:

In this section we are discussing the result of multinomial logistic regression (MLR). For child activities belonging to age group 4-9 years . the activity of child belongs to this age group are categorize into three categories (1) no schooling no work (NSNW) , (2) only schooling (OS).

Furthered the activity which is treated as dependent variable dependent variable takes the numeric child activity ‘0’ if activity is NSNW, ‘1’ if activity is OS and ‘2’ if activity is OW . For all model we have 19 independent variable including parenteral background information, household socioeconomic variable and demographic variable.

MLR is the estimated on both dataset (having 43302 observation out of 72562 observation) and testing dataset (having 29260 observation out of 72562 observation) in order to check the performance of the accuracy.

The results of MLR on training sample set result are provided Table no 4.1.

In multinomial logistic regression, one category of the dependent variable is taken as reference and separate model coefficients are estimated on the remaining levels. In our experiment child activities has three level. NSNW,OS,OW. By default the NSNW is taken as a reference. Therefore, for the remaining two activities OS and OW are get the model coefficients.

Probability equation.

The results of multinomial logistic regression are presented in table 5.1.according to these results being a urban region has positive impact on probability to going only school versus probability to activity of no schooling no work but result also show that it is insignificant variable at the 5% level. Our finding shows that ,in the urban areas are 0.019 time more likely to spend child activity in only school than in rural areas child. Empirical finding supported that the theoretically perspective ,that the those people live in city or the town they have more facility of education ,friendly environment, they have more parental support system ,they have more awareness about the value of education as compare to the people who live in the rural areas . Male gender have positive and significant impact on two probability of only school activity vs to the probability of no school no work activity. The odd ratio tell us the male child gender are 1.090 time more likely to involved only school activity as compare to the female child gender. Our empirical result supported that the male oriented Pakistani society .In this society, more dependence on the son than the daughters, parents are more concern about the male child education because they think that son can give better reward. In this model we used 4 to 9 child ages ,these child ages are positive and significant impact on the ratio of two probabilities the odd ratio is greater than 1,indicates that the more likely to prefer the only school activities over the no school no work activity. Number of infants in family have negative and insignificant variable impact on child activity. Odd ratio for number of infants is less than 1 indicate that the more likely to prefer the no school no work activity over the only school activity. this scenario see in the theoretically perspective that the number of infants mean that number of sibling .when the size of sibling increase also increase the financial load on the family budget so in the limited resources parents do not prefer to child involve in school activity .number of elders in family have positive and significant impact on child activity. odd ratio of that numbers of elder is greater than 1 it means that more prefer to school activity over the no school no work activity. The empirical result also support that real society scenario that is when number of elders increase in the family they can financially support to younger sibling on the way of education for the betterment of future. The number of males and females having age between 16 to 64 years use in the regression, in which male variable have negative and insignificant impact on the probability of only schooling versus probability of no school no work activity. Odd ratio of this variable is less than one indicated that if the numbers of males increases we would expected that they are more likely to prefer the no schooling no work activity and the numbers of female have positive and significant impact on the probability of only schooling versus probability of no school no work activity. Odd ratio is almost equal to one it means that the numbers of female between the age group 16 to 64 are more likely to prefer the only school activity.

The parental background of child is represented by variables parents education and employment status which plays an important role in the decision in child activities .First we discuss the results of parental education and then the results of parental employment status. The results show that educated mother and father have the positive and significant impact on the probability of only school activity versus the no school no work activity and odd ratio of mother educated indicates that 1.536 time more likely to prefer the only school child activity than the non-educated mother. Odd ratio of educated father is 1.054 which shows that educated father are more likely to prefer the only school activity than the non-educated father. Empirical results also supported the real world situation ,when parents are educated their preferences is more to child involved in the school going activity rather than the other kind of activities because they know that investment on human building is capital building as well. Employment status of parents tell us the capacity of parents to invest or nor to invest in the human capital. In analysis employment status have four categories, results shows that the mother paid employment ,self-employment ,unemployment and unpaid worker are positive and only mother unemployment are significant impact on the probability of only school activity versus the no school no work activity. Odd ratio of mother employee as compare to all other categories shows that mother employee is more likely to prefer the only school child activity over the no school no work activity. In case of father employment status the result shows that father paid employee ,self-employment , unpaid worker have negative and are significant impact on the probability of only school activity versus the no school no work activity. . Odd ratio of father employee as compare to other categories are more likely to prefer the only school child activity over the no school no work activity. In the demographic variables ,number of cattle's and area of agriculture land variables have negative impact and numbers of cattle's are significant impact on the probability of only school activity versus the no school no work activity. Odd ratio of number of cattle's variable is closer to zero mean that more prefer to child activity over the no school no work activity. Odd ratio of agriculture land is greater than 1.Finally we analyzed the impact of socioeconomic variables in the choice of child activities. Socio-economic factors are helpful in the measurements of indirect poverty facing the household members and also investigate the standard of living in terms of accessibility. Socio-economic factors includes ,one room house, personal house, kacaha house ,availability of toilet facility ,source of drinking water and fuel use for cooking. Empirical results indicate that the piped water as a source of drinking water has positive and significant impact on the probability of only school activity versus the no school no work activity. Odd ratio indicates that those household using piped water are more likely to prefer the only school activity. Results show that a household who lives in one room house have negative and significant impact on child activity and odd ratio also indicates that they are more likely to prefer no school no work activity over school only .empirical result also supports the theoretical point of view i.e., living in one room house is the indication of indirect poverty and it is difficult for them to meet their basic needs so they cannot support the school only activity. whereas, the odd ratio of personal house is near to 1 which indicates that they prefer only school activity as compare to the those individuals who do not live in personal house.

Availability of sewage system and fuel used for cooking are important measurement of basic facilities availabe to households.no toilet facility in the house having odd ratio indicates that they are 0.24 time more likely to prefer the only school child activity over the no school no activity. Other toilet facility variable are positive impact on the probability of only school activity versus the no school no work activity but insignificant variable. In Pakistan main source of fuel for cooking is gas and almost every households in urban area are using the gas because it is comparatively cheaper than the electricity and In most of the villages gas facility is not available therefore, house hold uses the other source of fuel like wood for cooking. The other fuel used for cooking variable is negative impact and insignificant variable. Odd ratio indicates that as compare to the gas facility the other fuel for cooking is 0.24 time less likely to prefer the only school child activity. Households who are not living kaccaha house it indicates that family has basic facility of life and their shows that they are 1.212 time more likely to prefer the child only school activity than the those who live in kaccha house.

TABLE 4. 1: RESULTS OF ONLY SCHOOLING VS NO SCHOOLING NO WORK

Explanatory Variable	**β Coefficient**	**P value and level of significance**	**Odd ratio**
Urban	0.019	0.64	1.019
Educated Father	0.720	0.000**	2.055
Male	0.737	0.000**	2.091
Child age	0.553	0.000**	1.738
Numbers of female between 16-64	0.022	0.18*	1.023
Numbers of Male between 16-64	-0.090	0.000**	0.913
Father Employment status			
Paid employment	-0.319	0.15***	0.726
Self - Employment	-0.400	0.002***	0.651
unpaid Worker	-0.019	0.97	0.981
Mother Employment status			
Paid employment	0.650	0.35	1.917
Self - Employment	0.929	0.18*	2.533
Unemployed	0.829	0.23	2.922
unpaid Worker	0.646	0.35	1.9808
Source of drinking water			
Piped water	0.182	0.000**	1.201
Toilet facility			
other toilet facility	0.219	0.000**	1.246
Fuel for cooking			

Explanatory Variable	β **Coefficient**	**P value and level of significance**	**Odd ratio**
other fuel used	-0.272	0.000**	0.761
Types of house			
Pakka house	0.794	0.000**	2.212
One room house			
Yes	-0.290	0.000**	0.748
Personal house			
Yes vs No	0.106	0.004**	1.111
Educated mother	-0.930	0.000**	2.536
Number of cattle	-0.003	0.000**	0.996
Agriculture land	0.000	0.99	1.000
Number of infants	-0.013	0.32	0.986
Number of elders	0.046	0.100	1.047
Constant	-4.951	0.000**	0.007

*1%,**5%,***10%

Probabilities results for OW category is represented on Table 4.2

Firstly we discussed about parental information parents education as well parents employment status is important in the decision of child activity in the model of probability of only work activity vs no schooling no work activity father education and mother education are the negative impact and father education are the insignificant impact but mother education are the significant impact on the ratio of two probability. odd ratio of the father education indicated that as compare to non -educated father only 0.09 time more likely to prefer the only work activity .next we see that the employment status ,that is the very much important for preference the child activities if family is strong financially parents are more focus on the school activity as compare to the other .the empirical results show that the mother ,self-employment ,unemployment and unpaid worker are the negative impact only mother paid employment are the positive and insignificant impact on the probability of only work activity versus the no school no work activity . Odd ratios of that variables indicated that by compare the mother paid employee to all other categories are more likely to prefer the only work child activity over the no school no work activity.in the case of father the result show that father paid employee are negative and other categories are father self-employment , unpaid worker are positive and father paid employee ,self-employment , unpaid worker are significant impact on the probability of only work activity versus the no school no work activity. . Odd ratios of that variables indicated that by compare the father paid employee to all other categories are less likely to prefer the only work child activity over the no school no work activity.

TABLE 4. 2: RESULTS OF ONLY WORK VS NO SCHOOLING NO WORK

Explanatory Variable	**B coefficient**	**P value and level of significance**	**Odd ratio**
Urban	-0.197	0.087	0.821
Educated Father	-0.008	0.099	0.991
Male	0.363	0.52	1.439
Child age	0.373	0.046**	1.452
Numbers of female between 16-64	0.318	0.31	1.375
Numbers of Male between 16-64	0.041	0.001**	1.042
Father employment status			
Paid employment	9.780	0.000**	1.68
Self - Employment	11.270	0.000**	78445.80
unpaid Worker	15.70	0.000**	6.620
Mother employment status			
Paid employment	0.030	0.97	1.031
Self – Employment	-0.253	0.77	0.776
Unemployed	-0.842	010	0.430
unpaid Worker	-0.152	0.79	0.858
Source of drinking water			
Piped water	0.812	0.24	2.253
Toilet facility			
other toilet facility	-0.003	1.00	1.004
Fuel for cooking			
other fuel used	-0.014	0.99	0.985
Types of house			
Pakka house	-0.316	0.62	0.728
One room house			
Yes	0.228	0.72	1.256
Personal house			
Yes	29.44	0.000**	6.148
Educated mother	-26.54	0.000**	0.000

Explanatory Variable	**B coefficient**	**P value and level of significance**	**Odd ratio**
Number of cattle	-0.203	0.99	0.815
Agriculture land	-0.003	0.92	0.996
Number of infants	-0.186	0.58	0.829
Number of elders	1.168	0.003	3.218
Constant	-49.82	0.000**	0.000

*1%,**5%,***10%

4.2.1 Confusion Matrix of MLR (4-9) Through Training Dataset

The performance of multi logistic regression model is checked by the confusion matrix, represented on table 5.3. The correct classification observations from the training data are on the diagonal of the confusion matrix. Accordingly, the overall accuracy of the training data is 72%.

TABLE 4. 3: CONFUSION MATRIX OF MLR (4-9)

Activities	NSNW	OS	OW
NSNW	12868	5450	7
OS	6561	18487	6
OW	0	0	0

4.3 Classification Tree

Classification tree is supervised machine learning technique it is used to predict the qualitative outcome having continuous and categorical predictors. In this study as, our aim is to predict the three level of child activities based on both categorical and continuous features .classification tree predict the each observation that fit to the most commonly occurring class of training observation in t region to which class its belong .We have estimated the classification tree for child activity belonging to age group 4-9 in the R software. the results are represented in table 4.4 and decision tree is presented in figure.

TABLE 4. 4: COMPLEXITY TABLE FOR CHILD ACTIVITY BETWEEN AGE 4-9

Variable used in the construction of tree; Child age, father education ,Child gender, kaccha house, mother education ,one room house, toilet facility.				
Root node Error : 19472/43504 = 0.44572				
Sample : 43504				
CP	N split	Rel .error	x. error	X std
0.265	0	1.000	1.000	0.005
0.028	1	0.734	0.734	0.005
0.019	3	0.676	0.676	0.004
0.016	4	0.656	0.660	0.004
0.016	5	0.640	0.643	0.004
0.008	6	0.623	0.623	0.004
0.002	7	0.614	0.614	0.004
0.002	12	0.610	0.615	0.004
0.001	15	0.6087	0.6123	0.004

According to classification tree child’s age, father education, child’s gender, kaccha house ,mother education ,one room house information and toilet facility variables are the main features of child activity. It can also be seen that in the tree diagram the splits are occurred only these variables. root node error tell us the error rate of single node and n is the number of training observation. Total sample of child activity model having age between 4-9 years is consist of 72562 observation, which is further divided into 60% training (i.e.43504) and 40% testing sample. The complexity Table 5.4 provides the information on complexity parameter(cp), number of split (nsplit),cross validation error rate (xerror) ,standard error(xstd) and resubstitution error rate (rel error).Complexity parameter is the user define parameter, it prune off the unnecessary splits and save the computational time. Number of split indicates that total 15 splits are occurred and also visible in tree diagram. Resubstitution error rate is the amount of original observation that are misclassification by the serval subset of the original tree. This rate is minimized when the tree grow . Larger tree having lowest resubstition error rate so the choosing of tree which have lowest rate is not the best choice so we used cross validation as the alternative of resubstition error rate for the selecting the tree.

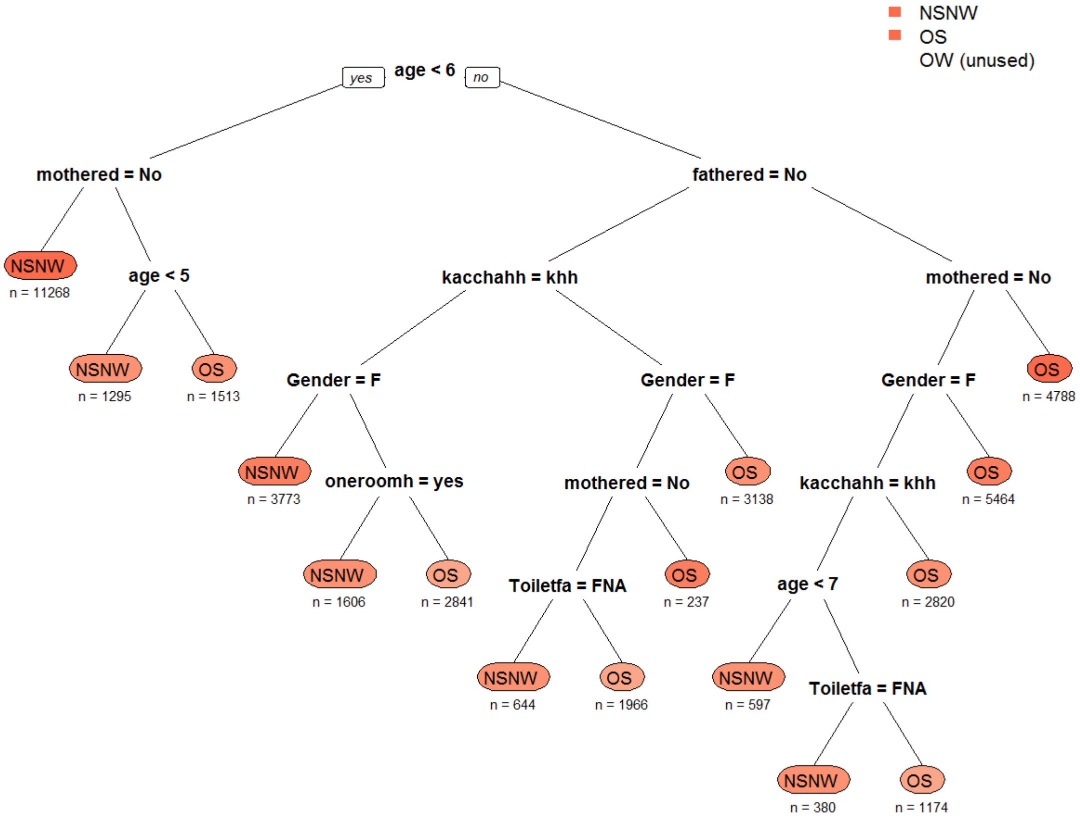
4.3.1 Classification Tree for Child Age Group Between 4-9

Classification tree shows that the age is the most important variable after the mother education and father education in the determines of child activity. Hence ,age is the root node where classification tree can grow from this point .left branch of the tree indicates that if child age is less than 6 year and having uneducated mother then predicted than child can involve in no schooling no work activity group. similarly, move to the next variable again i.e., if age is less than 5 years then the predicted is child involved in no schooling no work activity. On the other hand if age is greater than 5 years then the predicted child is involved in only schooling activity.

On the right branch of the tree ,if father and mother both are educated then we predicted that the child can involve in only schooling activity . likewise, if mother is non educated and child gender is male then we predicted that child in only schooling activity and if child gender is female ,and living in pakka house then we predicted that child is involved in only schooling activity, if child age is greater than 7years and they have toilet facility in the house then predicted child involved in only schooling activity. If mother is non educated and child belongs to the females gender, living in the kaccha house and child age is less than 7 year then we may predict that child is involved in the no schooling no work activity group.

In Middle branch of the tree, if father is non-educated they live in the kaacha house and child gender is female then child belongs to no schooling no work activity .if child belongs to male gender and they live in the more than one room house then predicted that child involved in the only school activity .if child family live in the kachha house, belongs to female gender ,mother is non educated and toilet facility is not available in the house then predicted that child involved in the no schooling no work activity and if father is non educated and they live in the pakka house, child belongs to male gender then we predicted that child involved in the only school activity.

FIGURE 4. 1: DECISION TREE; CHILD ACTIVITY FOR AGE 4-9



4.3.2 Confusion Matrix of CT Through Training Dataset

According to confusion matrix the overall accuracy ratio of training sample under classification tree is 73% and misclassification rate is 26.4%.for testing sample the accuracy rate is slightly increase to 73.57% and misclassification rate is reduced to 26.38%.hence, the accuracy rate of both training and testing sample are similar to each other.so it is clearly evidence that there is no problem of overfit.

TABLE 4. 5; CONFUSION MATRIX OF CT

Activities	NSNW	OS	OW
NSNW	13675	5880	8
OS	5786	18152	3
OW	0	0	0

4.4 Random Forest

Random forest is an ensemble method. It can provide better prediction and improve the results by building multiple decision tree and merge them together. We have choose n 200 numbers of trees randomly. If the number of tree increases then there no effect on the error rate. This is depicted in the plot of random forest ,figure .it indicates that error for different classes and the out of bag (OOB) sample over the number of tree are equal to 26.02% is in black pattern and other are in different color such that error of NSNW activity is in red color, error of OS in green color and error of OW is in blue color.

4.4.1 Error plot of random forest

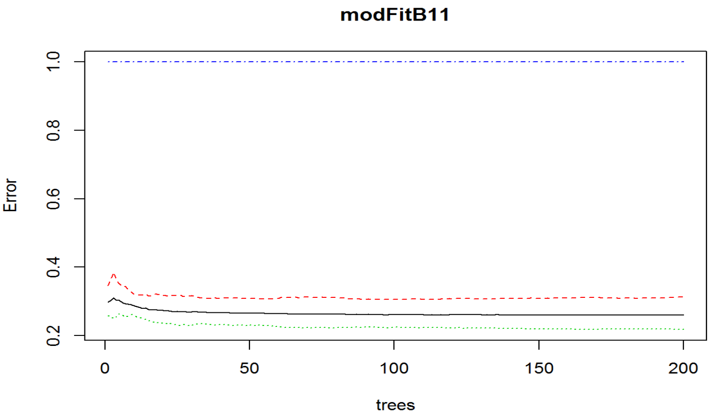


FIGURE 4. 2: ERROR OF RANDOM FOREST FOR DIFFERENT CLASSES USING TRAINING DATA.

TABLE 4. 6: RESULTS OF RANDOM FOREST

Types of Random forest : Classification				
Number of Trees(n) = 200				
No. of variable tried at each split: 2				
OOB (error rate) :26.02%				
Confusion Matrix of training data				
Predicted	NSNW	OS	OW	Class. Error
NSNW	13324	6047	0	0.312617
OS	5247	18818	0	0.218034
OW	5	6	0	1.000000

In this output confusion matrix is represented for training dataset .in which we can see that 30.66% correct prediction that are belongs from the NSNW group ,43.31% correct prediction that are belongs from the OS group and we can observed that no correct prediction come from the OW groups .the class. Error of group 1(NSNW) is 31.26%, group 2(OS) is 21.8% and the group 3(OW) is 100 %

However ,we are more interested to know which variable is more important in the building of child activity model .two measures, (i) mean decrease accuracy and (ii)mean decrease Gini are commonly used in the investigation of important variables under random forest. Figure 5.3 is represent the plot of important variables.

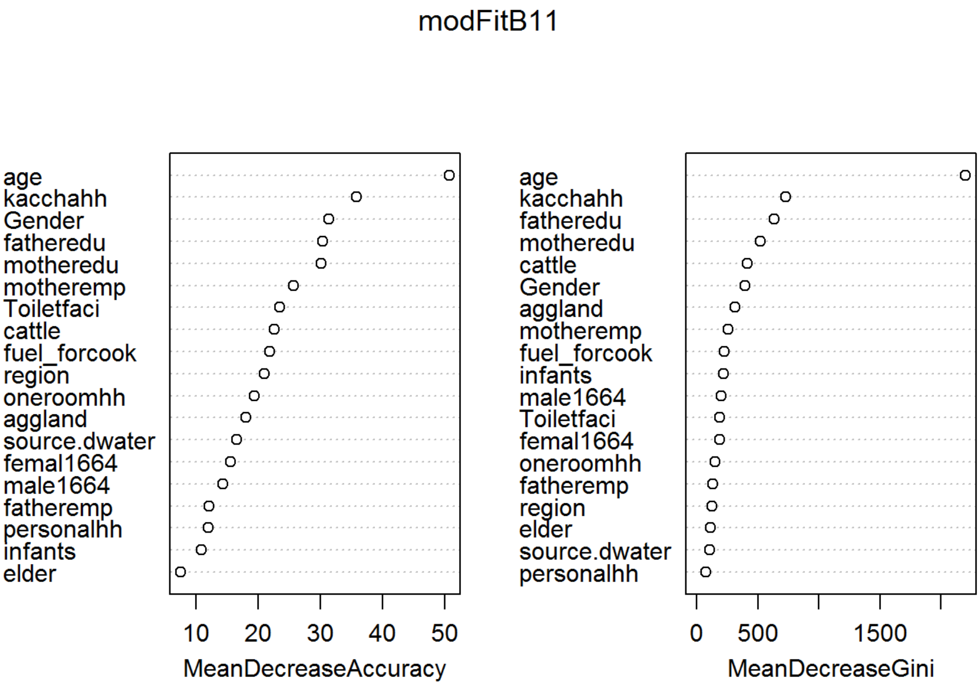


FIGURE 4. 3: PLOT OF IMPORTANT VARIABLE

Accordingly ,both measures of plot of important variable show that child’s age is most important variable,then condition of house(kaccha house)and rest of other variable.

4.4.2 Prediction from Random Forest by Using Testing Dataset

In the table 4.7,confusion matrix is represented for test dataset .in which table off diagonal represent miss classification observation and on diagonal represent correct classification , in which we can see that 0.328 or 32% correct prediction that are belongs from the NSNW group , 0.458 or 45% correct prediction that are belongs from the OS group and we can observed that no correct prediction come from the OW groups .total accuracy of model is 0.7874 or 78% .

TABLE 4. 7: CONFUSION MATRIX OF RF BY USING TEST DATA

Activities	NSNW	OS	OW
NSNW	9563	2689	2
OS	3491	13363	7
OW	0	0	0

4.5 Linear Discriminate Analysis for Child Age Between 4-9

LDA is another technique of data classification .it assumes that all variables must be normally distributed with in each group. However, real world data are rarely normally distributed and may lead to effect the performance of LDA.

We have performed the shapiro-wilk (1965) test on our child activity data to test the null hypothesis of population under consideration is normally distributed. For all variables the p-value of shapiro-wilk test is less than 0.05 ,therefore, we reject the null hypothesis and conclude that child activity data is not multivariate normally distributed.

The another main assumption of LDA is to have same covariance among dependent and independent variables. This assumption is investigated through the Box M test .The null hypothesis of this test is that covariances matrix are equal cross each group. Our results show that the p value of box M test is 0.000 therefore, we reject the null hypothesis and conclude that covariances matrix are not equal cross the each group. Regardless the violation of assumption we have performed LDA on child activity data for solve of empirical analysis and comparison of LDA with the earlier classification techniques.

4.5.1 Result of Linear Discriminate Analysis

The prior probability of child activities under the LDA which are categorize in three group are π_1=0.449269 for NSNW , $\pi_2 = 0.550362$ for OS and π_3=0.000368 for OW activity.it means that 44% of the training observations are belongs to no schooling no work activity ,55% observations are belong to only school activity and 0.036% observations are belong to only work activity .

This coefficients of linear discriminate analysis give us the linear combination of all variables which are used as the form of LDA decision rule In other words, these are the multipliers of the elements of predictor variables. If -0.04144 $* regionU + 0.0606972 * sexM + 0.48038 * age + 0.012168 * infants + 0.03526 * elder - 0.07050 * male16 - 64 + 0.27169 * female16 - 64 - 0.17251 * Fatherpaidemp - 0.23852 * Fatherselfemp + 0.058718 * fatherunpaidwork + 0.481422 * motherpaidemp + 0.68584 * motherselfemp + 0.619128 * motherunemp + 0.460480 * motherunpaidwork + 0.18700 * pipedw + 0.229318 * OTF - 0.12094 * OFU + 0.17156 * NKH - 0.2570 * oneroomyes + 0.023 * personalhhyes - 0.002529 * cattle + 0.002175 * aggland + 0.6753 * motheredu + 0.6375 * fatheredu$ is larger than Lda is classify that they belongs to no schooling no work(NSNW) activity and if is less than lda is classify that they not belongs to no schooling no work(NSNW) activity they belong to OW activity. Similarly, if .if -0.90505 $* regionU + 0.27512 * sexM + 0.12636 * age + 0.08370 * infants + 1.21002 * elder + 0.56698 * male1664 - 0.305171 * female1664 + 0.17382 * Fatherpaidemp + 0.927085 * Fatherselfemp - 0.515194 * fatherunpaidwork + 0.68778 * motherpaidemp + 0.31089 * motherselfemp + 0.37863 * motherunemp + 0.2873 * motherunpaidwork +$

$0.34284 * pipedw - 0.23991 * OTF\ - 0.09351 * OFU - 0.29275 * NKH + 0.67095 * oneroomyes + 0.98596 * personalhhyes + 0.020122 * cattle - 0.00329 * aggland - 0.290131 * motheredu - 0.05558 * fatheredu$ is larger than Lda is classify that they belongs to only schooling (OS) activity if it is less than they belongs to only working activity .

4.5.2 Plot of LDA Model

Linear discriminant plots are obtained through LDA function for each training observation .in these plots show that how LDA classify the group with the histogram. These plots also show that the separation of three groups along with the overlying areas ,these overlying or error are possible when they predicting the groups.

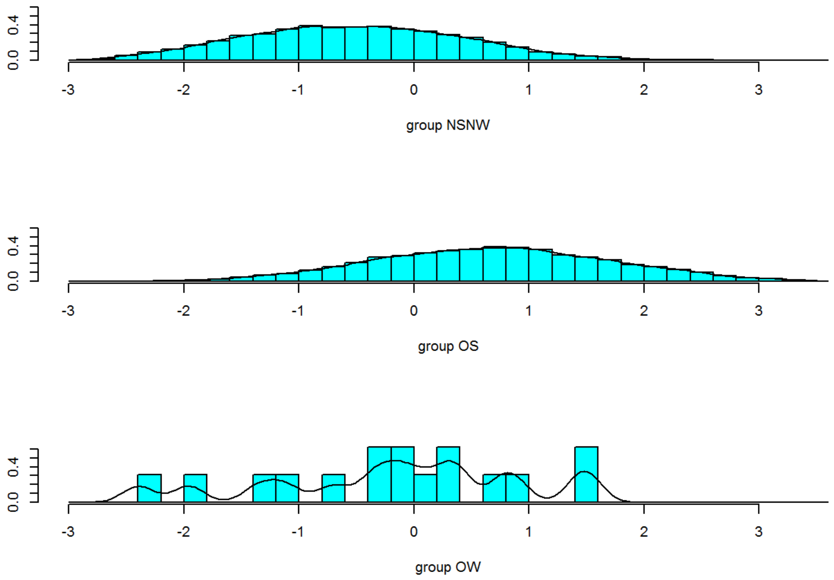


FIGURE 4. 4: PLOT OF LDA

4.5.3 Confusion Matrix by Using the Training Dataset

Confusion matrix represented table 4.8, that on diagonal have correct classification and off diagonal have miss-classification training observations .it can also observed that 30.04% correct prediction that are belong from group no schooling no work activity and 42.01% prediction are belong from the group only schooling activity. That indicates the overall accuracy of training model is 72% .

TABLE 4. 8: CONFUSION MATRIX OF LDA BY USING TRAINING DATASET

Predicted	NSNW	OS	OW
NSNW	13055	5656	8
OS	6467	18258	8
OW	1	2	0

4.6 Evaluating the Above Techniques on the Basis of Confusion Matrix and ROC Analysis

In this section we are evaluating the accuracy multinomial logistic regression(MLR),Classification and regression tree (CART) and Linear discriminant analysis (LDA) on the basis of two methods that is (i)accuracy rate based on confusion matrix and (ii) ROC analysis .

4.6.1 Accuracy Based on Confusion Matrix for All Techniques

TABLE 4. 9; ACCURACY RATES OF CONFUSION MATRIX FOR ALL TECHNIQUES

Techniques	Testing dataset
MLR	71%
CT	73%
LDA	72%

4.6.2 Accuracy Measurement Based on ROC Analysis

The ROC analysis of MLR, CT and LDA are presented in figure 4.5,4.6, and 4.7. In figure 4.5 represent the Multi ROC curve of MLR, The middle line show that the area under the curve and other two curve show that specificity and sensitivity of each group of activity .area under the curve is actual measure of accuracy .the AUC of this model is 0.6451 .

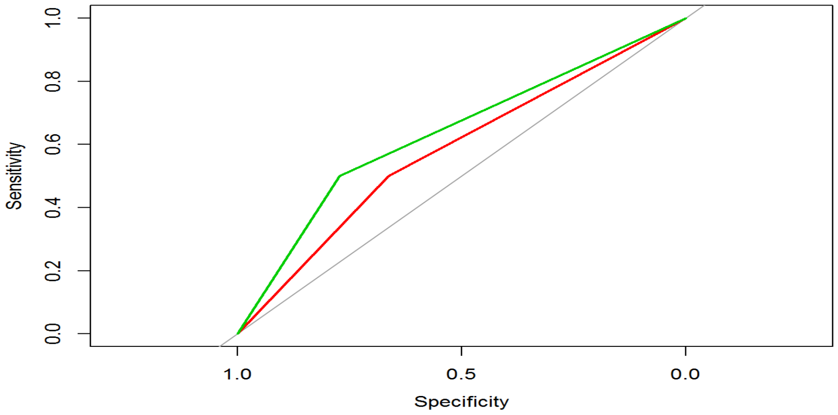


FIGURE 4. 5: MULTI-ROC ANALYSIS OF MLR MODEL

in the figure 5.6 show Multi ROC curve ,in this curve give us the sensitivity and specificity of each groups activity and multi-class area under the curve .0.702 sensitivity ,0.755 specificity of class no schooling no work(NSNW).0.755 sensitivity and 0.702 specificity of class only schooling(OS) and 0.000 sensitivity ,1.00 is specificity of class only work(OW). The area under the curve is 0.652 actually that is the accuracy of classification tree model

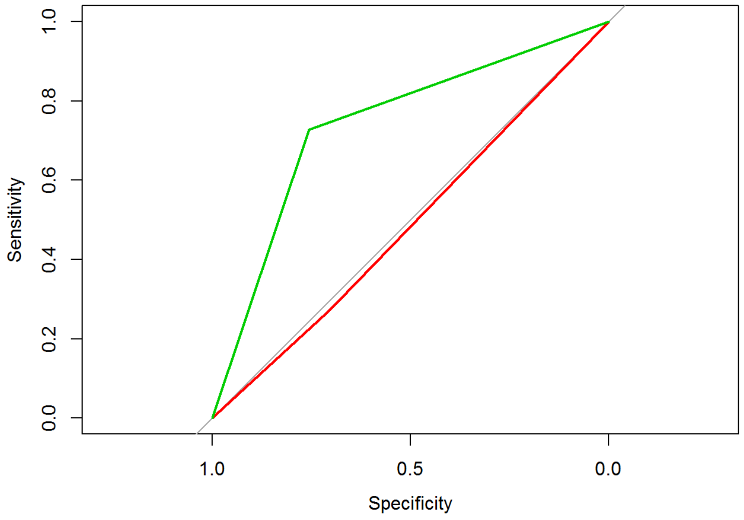


FIGURE 4. 6: MULTI- ROC CURVE OF CLASSIFICATION TREE

In the figure 4.7 ,The middle line show that the area under the curve and other two curve show that specificity and sensitivity of each group of activity .area under the curve is actual measure of accuracy .the AUC of that model is 0.6441

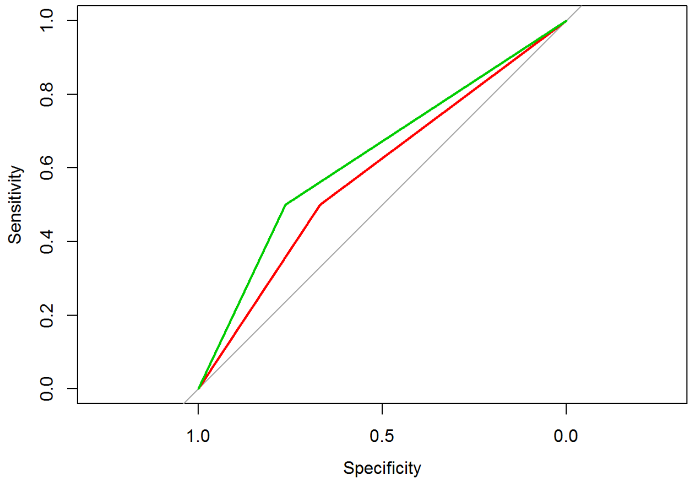


Figure 4. 7: Multi- Roc curve of LDA

Our results indicated that when we compare the three techniques for child activities between the age group 4-9 ,conclude that the machine learning technique (classification tree method) is better than multinomial logistic regression and linear discriminate analysis on the basis of both method.

4.6 Multinomial Logistic Regression for Child Activities Between age 10-14

In this section we are discussing the results of multinomial logistic regression(MLR) for child activities belong to age group 10-14 years. The total sample consist of 53280 observations which is divided into two non-overlapping datasets, training and testing dataset, training dataset have 32106 observations and testing dataset have 21174 observations. The activities of child belong this age group are categorize into four categories (1) no schooling no work (NSNW), (2) only schooling ,(3)only work ,(4) work with schooling.

Multinomial logistic regression was run to test that which variable is more effective in the decision of children choosing the any (four)child activities. The dependent variable is converted in to polychotomous variable i.e. child activity =0 for no schooling no work ,child activity =1 for only schooling , child activity =2 for only work, child activity =3 for work with schooling .No schooling no work is also the reference category therefore, the coefficient ,level of significant and odd ratio of other categories will be obtained with respect to reference category .The results of MLR are shown Table

According to these results belonging to urban region has positive and significant impact on the probability of only schooling verses the probability of no schooling no work . Odd ratio tells that child living in urban region are 0.165 time more likely to have only school activity as compare to child living in rural area . Empirical finding also supports the ground reality ,that those live in city have more facilities of education and well aware about the value of education as compare to those who live in the rural areas .

Male gender has positive and significant impact on the probability of only school activity versus to the probability of no school no work activity. The odd ratio represents that male child are 3.499 time more likely to involved in only school activity as compare to the female child . it represents male oriented Pakistani society .in male oriented society, dependence on the son is more than the daughters, therefore, parents are more concern about the male child education because they think that son can give better reward.

In this model we used 10 to 14 child ages ,these child ages are positive and significant impact on the ratio of two probabilities , odd ratio is less than one indicates that the more likely to prefer the no schooling no work activities.

Numbers of infants in family have negative but significant variable impact on child activities. Odd is less than one which indicates that it is more likely to prefer the no school no work activity over the only school activity. number of infants mean that numbers of sibling .when the size of sibling increase lead to also increase the financial load on the family budget so in the limited resources parents do not prefer to child involve in school activity .numbers of elder in family have positive and significant impact on child activities, odd ratio of that variable is greater than 1 it means that more prefer to school activity over the no school no work activity. The empirical result also support that real society scenario that is when numbers of elder increases in the family they can financially support to younger sibling on the way of education for the betterment of future.

The number of males and female having age between 16 to 64 years use have negative but significant impact on the probability of only schooling versus probability of no school no work activity. Odd ratio of this variable is less than one indicates that if the numbers of male increases we would expected that they are more likely to prefer the no schooling no work activity. Whereas, numbers of female have positive and significant impact on the probability of only schooling versus probability of no school no work activity. Odd ratio is greater than one, it means that the numbers of female between the age group 16 to 64 are more likely to prefer the only school activity.

The parental background of child is represented by parents education. it plays an important role in the decision of child activities. The results show that educated mother and father have the positive and significant impact on the probability of only school activity versus the no school no work activity and odd ratio of educated mother indicates that 2.367 time more likely to prefer the only school child activity than the uneducated mother. Odd ratio of educated father is 1.516 which shows that educated father are more likely to prefer the only school activity than the uneducated father. Empirical results support the real world situation ,educated parents prefer to engage their children in the school going activity rather than the other kind of activities because they know that investment on human building is capital building as well.

Next we discuss the significant of demographic characteristics , Number of cattle has significant impact on the probability of only school activity versus the no school no work activity. Odd ratio of this variable is greater than one mean that more prefer to only schooling activity.

Finally we have analyzes the impact of socioeconomic variables in the choice for child activities Socioeconomic factors are helpful in the measurements of indirect poverty facing the household members and also represent the standard of living in terms of accessibility. socioeconomic factors includes one room house, personal house, kacaha house ,availability of toilet facility ,source of drinking water and fuel use for cooking. Empirical results show that piped water as a source of drinking water has positive and significant impact on the probability of only school activity versus the no school no work activity. Odd ratio indicates that those household using piped water are more likely to prefer the only school activity.

Results show that a household who lives in one room house negative and significant impact on child activity and odd ratio also indicates that they are more likely to prefer no school no work activity . Empirical result also supported the theoretical point of view that there who are living in the one room houses are suffering from poverty and may be difficult to hand the basic necessary so they cannot afford the school only activity. Odd ratio of personal house is near to 1 that indicates the those who live in the personal house are likely to prefer only school activity as compare to the those individual who do not live in the personal house.

Availability of sewage system and fuel used for cooking is important indicators to measure the availability of basis facilities. In the household. In Pakistan main source of fuel for cooking is gas and almost every households in urban area are using the gas because it is comparatively cheaper than the electricity. In most of the villages gas facility is not available, therefore ,house hold uses the other source of fuel like wood for cooking .The other fuel used for cooking variable has negative and significant impact. Its odd ratio indicates that as compare to the gas facility the other fuel for cooking is 0.432 time less likely to prefer the only school child activity . Households who are not living in kaccaha house it indicates that family has basic facility of life. And their odd ratio shows that 0.9 time more likely to prefer the child only school activity than the those who live in kaccha house.

TABLE 4. 10: RESULTS OF ONLY SCHOOLING VS NO SCHOOLING NO WORK

Explanatory Variable	**B coefficient**	**P value and level of significance**	**Odd ratio**
Urban	0.152	0.003**	1.165
Educated Father	0.922	0.000**	2.516
Male	1.503	0.000**	4.499
Child age	-0.177	0.000**	0.837
Numbers of female between 16-64	0.079	0.000**	1.082
Numbers of Male between 16-64	-0.087	0.000**	0.916
Father employment status			
Paid employment	-0.302	0.081	0.739
Self - Employment	-0.380	0.027	0.683
unpaid Worker	0.332	0.529	1.395
Mother employment status			

Explanatory Variable	**B coefficient**	**P value and level of significance**	**Odd ratio**
Paid employment	0.285	0.693	1.330
Self – Employment	0.744	0.303	2.105
Unemployed	0.431	0.550	1.539
unpaid Worker	0.480	-0.506	1.616
Source of drinking water			
Piped water	-0.331	0.000**	1.392
Toilet facility			
other toilet facility	-0.336	0.000**	0.713
Fuel for cooking			
other fuel used	-0.565	0.000**	0.568
Types of house			
Pakka house	0.645	0.000**	1.908
One room house			
Yes	-0.338	0.000**	0.713
Personal house			
Yes	0.072	0.148	1.074
Educated mother	1.196	0.000**	3.367
Number of cattle	-0.002	0.040**	0.997
Agriculture land	0.001	0.115	1.001
Number of infants	-0.122	0.000**	0.884
Number of elders	0.057	0.096	1.059
Constant	1.660	0.025**	5.262

*1%,**5%,***10%

According to table these results show that father education has negative and significant impact on the probability of only work activity versus the no schooling no work activity ,odd ratio of this variable indicates that 0.17 time less likely to prefer the only work activity than the non-educated father. Male gender have positive and significant impact on the probability of only work verses the no schooling no work activity ,the odd ratio tells that male child gender are 2.515 time more likely to involved only work activity as compare to the female child gender .our empirical result are support the situation when family is in financial cries then family burden is on the shoulder of male child and family push the male child for the paid work . Pushing on the male child as compare to the female for work due to more market opportunities and parents are more concern about female safety issues . The numbers of male between the age group 16-64 years has the negative significant impact on probability of only work verses the no schooling no work activity, odd ratio is less than 1 indicated that if the numbers of males increase we would expected that they are more likely to prefer the no schooling no work activity similar results are obtain for the numbers of female between the age group 16-64 years are negative but significant impact on the probabilities of only work verses no schooling no work activity and odd ratio is less than 1 indicates that if the numbers of females is increase than more likely to prefer the no schooling no work activity. child’s age can play an important role in the decision of activity our results show that child age has positive and significant impact on the ratio of probability of only work verses no schooling no work and odd ratio is less than 1 its mean that more likely to prefer the no schooling no work activity.

Next we discuss the impact of socioeconomic variables in the choice of child activities. Socioeconomic factors are helpful in the measurements of indirect poverty facing the household members and also investigate the standard of living in the term of accessibility. our empirical result indicates that the piped water used as the source of drinking water has positive and significant impact on the ratio of only work versus no schooling no work ,odd ratio of that variable indicates that 0.43 time less likely to prefer the only work activity as compare to the other source of drinking water . Availability of sewage system are the measurement of basic facilities available to the household. The coefficient of other toilet facility variable is positive and have significant impact on two ratio and it's odd ratio indicates that theses household more likely to prefer the only work activity as compare to the household where toilet facility is not available. Our empirical results supported the real word situation if socioeconomic variable is good or availability of basic resources then child cannot more involved in only work activity

Number of cattle's in the house variable is positive and have significant impact on the probability of only work versus no schooling no work activities. Odd ratio is greater than 1 its indicates that child are involved in the only work activity . our result suggest that cattle increase in the house then children are engage to take care the animals.

Similar numbers of infants and number of elder in the family both variable have positive and significant impact on the probability of only work versus no schooling no work . Odd ratio of these variables is greater than 1 it means that more prefer to only work activity .this scenario see in the theoretically perspective that the number of infants mean that the number of siblings ,when the number of siblings and elder increase in the house also increase the financial load on the family budget therefore ,whit the limited resources of parents would prefer to involved the child in only work activity.

TABLE 4. 11: RESULTS OF ONLY WORK VS NO SCHOOLING NO WORK

Explanatory Variable	B coefficient	P value and level of significance	Odd ratio
Urban	0.135	0.25	1.145
Educated Father	-0.185	0.001**	0.830
Male	1.257	0.000**	3.515
Child age	0.438	0.000**	1.550
Numbers of female between 16-64	-0.314	0.000**	0.871
Numbers of Male between 16-64	-0.179	0.000**	0.835
Father employment status			
Paid employment	-0.361	0.342	0.696
Self - Employment	0.128	0.733	1.137
unpaid Worker	0.340	0.728	1.406
Mother employment status			
Paid employment	0.185	0.871	1.204
Self – Employment	-0.028	0.986	0.972
Unemployed	-1.177	0.302	0.308
unpaid Worker	0.310	0.785	1.364
Source of drinking water			
piped water	-0.565	0.000**	0.568
Toilet facility			
other toilet facility	0.160	0.003**	1.174
Fuel for cooking			
other fuel used	0.030	0.810	1.031
Types of house			

Pakka house	0.000	0.986	1.001
One room house			
Yes	-0.104	0.074	0.900
Personal house			
Yes	-0.177	0.030**	0.837
Educated mother	-0.104	0.458	
Number of cattle	0.006	0.000**	1.006
Agriculture land	-0.003	0.898	0.999
Number of infants	0.123	0.000**	1.131
Number of elders	0.176	0.002**	1.193
Constant	-6.367	0.000**	0.001

*1%,**5%,***10%

According to the results most of variables are insignificant, but some independent variables are significant and have strong impact on the probability of work with schooling versus no schooling no work activities . First we discuss effect of parental education ,results show that educated father and educated mother both have positive and significant impact on the probabilities of work with schooling versus no schooling no work activities. Odd ratio of indicates that as compare to non-educated father the educated father 0.41 time more likely prefer the work with schooling activity. Similarly odd ratio of educated mothers indicated that 0.86 time more likely to prefer the work with schooling activity as compare to the uneducated mother. parental education can plays very important role in the decision of child activity. If both mother and father are educated their preferences is more to child involved to the schooling going activity because they know that investment on human building is actually the building of capital as well .

Our finding show that child belong to urban region variable is positive and significant impact on the ratio of two probabilities. The odd ratio indictes that 0.53 time more likely spend time on both work and schooling activities than the rural areas child ,empirical finding also supported the theoretically perspective that those people live in the city they have more facility of part time job ,they have more parental support to purse the both activities so they do paid work along with the study because they are well aware about the value of education as compare to the people who live in the rural areas.

Child being Male has positive and significant impact on the probabilities of work with schooling versus no schooling no work activities, odd ratio tell us 10.09 time more likely to prefer the both activity as compare to the female child. Numbers of male between age group 16- 64 have positive and significant impact on the probabilities of work with schooling versus no schooling no work activities, odd ratio is less than one indicated that more likely to prefer the no schooling no work activity.

TABLE 4. 12: RESULTS OF WORK WITH SCHOOLING VS NO SCHOOLING NO WORK

Explanatory Variable	**B coefficient**	**P value and level of significance**	**Odd ratio**
Urban	0.616	0.001**	1.853
Educated Father	0.343	0.001**	1.410
Male	2.406	0.000**	011.09
Child age	0.308	0.000**	1.361
Numbers of female between 16-64	-0.033	0.564	0.966
Numbers of Male between 16-64	-0.131	0.011**	0.876
Father employment status			
Paid employment	-2.594	0.037	0.074
Self - Employment	0.613	0.421	1.846

unpaid Worker	1.161	0.245	5.053
Mother employment status			
Paid employment	-2.594	0.037**	0.074
Self – Employment	-1.559	0.203	0.210
Unemployed	-2.759	0.023	0.063
unpaid Worker	-1.213	0.319	0.297
Source of drinking water			
piped water	-0.074	0.636	0.927
Toilet facility			
other toilet facility	-0.203	0.081	0.815
Fuel for cooking			
other fuel used	-.234	0.273	0.790
Types of house			
Paka house	0.075	0.476	1.079
One room house			
Yes	-0.146	0.220	0.864
Personal house			
Yes	0.093	0.597	1.098
Educated mother	0.624	0.001**	1.867
Number of cattle	0.003	0.233	1.003
Agriculture land	0.004	0.140	1.004
Number of infants	0.115	0.045**	1.122
Number of elders	0.118	0.290	1.125
Constant	-6.368	0.000	0.001

*1%,**5%,***10%

4.6.1 Confusion Matrix of MLR(10-14) through Training Dataset

Table 5.11,represents the prediction through confusion matrix on the training dataset . We observed that 14.83%predicted response belongs to NSNW activity , 54.14% of predicted response belong to OS activity , 0.87% of predicted response belongs from OW activity and 0 %predicted response belongs from WWS activity whereas ,total accuracy of MLR model by using training dataset is 69.86 it means that the overall error of MLR model is 30.14%.

TABLE 4. 13: CONFUSION MATRIX OF MLR FOR TRAINING DATASET

predicted	NSNW	OS	OW	WWS
NSNW	4764	2263	793	60
OS	4602	17384	1152	340
OW	160	252	282	54
WWS	0	0	0	0

4.7 Classification Tree for child Age Between 10-14

Construct the decision tree for child to age group 10-14 activities are classified into four categories, no schooling no work (NSNW),only schooling(OS),only work (OW) and work with schooling (WWS) activities. The result of classification tree on the training dataset and results are shown below.

TABLE 4. 14: SUMMARY OF CLASSIFICATION TREE

Variables actually used in the construction of tree or significant variables Father education, child gender, kaccha house, fuel for cooking ,mother education ,mother employed, region, child age, infants ,Toilet facility ,agriculture land,cattle, no of 16-64female,source of drinking water ,father employment	
Total sample	31955
Number of terminal nodes	23
Misclassification error rate	29.91

Summary of classification tree shows that total observation of training dataset is 31955 in which 15 variables are significant or important and most important variable is father education ,second one is child gender and then other variables. Total numbers of terminal/decision nodes are 23 ,misclassification error is 29.91.

TABLE 4. 15: INFORMATION OF INTERNAL NODES

Node	**Split point**	**N**	**Loss**	**Y value**	**Y probability**
1	Root	31955	12227	OS	0.2993,0.6173,0.0692,0.0140
2	Father education(NO)	15861	8425	OS	0.4077,0.4688,0.1065,0.0168
4	Gender (female)	7193	2995	NSNW	(0.5836,0.3319,0.0797,0.0045)
8	Kaccha HH (kaacha hh)	3739	1133	NSNW	(0.6969,0.2115,0.0861,0.0053)*
9	Kaccha hh (NO kaacha hh)	3454	1857	OS	(0.4609,0.4623,0.0729,0.0037)
18	Fuel for cooking (other fuel)	2785	1395	NSNW	(0.4983,0.4096,0.0883,0.0003)
36	Mother education (NO)	2613	1268	NSNW	(0.5147,0.3892,0.0922,0.0038)
72	Mother employment (employee, U.E)	1463	640	NSNW	(0.5625,0.4203,0.0157,0.0013)
144	Region (Rural)	1227	529	NSNW	(0.5857,0.3946,0.0180,0.0015)*
145	Region (Urban)	186	75	OS	(0.4032,0.5967,0.0000,0.0000)*

73	Mother emp(paid.emp,self.emp,unpaid)	1150	628	NSNW	(0.4539,0.3495,0.1895,0.0069)
146	Child age (< 12.5)	715	406	NSNW	(0.4321,0.4237,0.1356,0.0083)
292	Infants (>=0.5)	267	138	NSNW	(0.4831,0.3333,0.1647,0.0187)*
293	Infants (< 0.5)	448	234	OS	(0.4017,0.4776,0.1183,0.0022)
586	Toilet facility (other facilities)	150	80	NSNW	(0.4666,0.3666,0.1600,0.0066)*
587	Toilet facility (facility not available)	298	139	OS	(0.3691,0.5335,0.0973,0.0000)*
147	Child age (>=12.5)	435	222	NSNW	(0.4896,0.2275,0.2781,0.0045)*
37	Mother education (Yes)	172	48	OS	(0.2500,0.7209,0.0290,0.0000)*
19	Fuel for cooking (Gas)	669	213	OS	(0.3049,0.6816,0.0089,0.0044)*
5	Gender (male)	8668	3620	OS	(0.2617,0.5823,0.1287,0.0271)*
10	Kaccha hh (kaccha hh)	4669	2312	OS	(0.3069,0.5048,0.1599,0.0282)*
20	Agriculture land (< 0.05)	3021	1655	OS	(0.3574,0.4521,0.1655,0.0248)
40	Mother employment (self.emp,U.E)	2123	1046	OS	(0.3537,0.5073,0.1205,0.0183)*
41	Mother employment(p.emp,unpaid work	898	569	NSNW	(0.3663,0.3218,0.2229,0.0405)*
164	Cattle (< 4.5)	488	265	NSNW	(0.4569,0.3565,0.1577,0.0286)*
165	Cattle (>=4.5)	252	164	OW	(0.2777,0.3095,0.3492,0.0634)*
330	Females 16-64 (>=1.5)	81	47	NSNW	(0.4197,0.3333,0.1851,0.0617)*
331	Females 16-64 (< 1.5)	171	98	OW	(0.2105,0.2982,0.4269,0.0643)*
83	Child age (>=13.5)	158	79	OW	(0.2278,0.2341,0.5000,0.0379)*
21	Agricultural land (>=0.05)	1648	657	OS	(0.2141,0.6013,0.1498,0.0345)*
11	Kachha hh (No kaccha hh)	3999	1308	OS	(0.2090,0.6729,0.0922,0.0257)*
3	Father education (yes)	16094	3802	OS	(0.1925,0.7637,0.0324,0.0112)
6	Gender (female)	7504	2507	OS	(0.29999,0.6659,0.029,0.0050)
12	Mother education (No)	5171	2299	OS	(0.3995,0.5554,0.0398,0.0052)
24	Kaccha hh (kaccha hh)	1906	901	NSNW	(0.5272,0.4071,0.0582,0.0073)
48	Child age (>=11.5)	1111	471	NSNW	(0.5760,0.3348,0.0810,0.0081)*
49	Child age (< 11.5)	795	391	OS	(0.4591,0.5081,0.0264,0.0062)
98	Father employment (self. Employed)	425	203	NSNW	(0.5223,0.4305,0.0376,0.0094)
196	Source of drinking (other source water)	355	161	NSNW	(0.5464,0.3971,0.0450,0.0112)*
197	Source of drinking (piped. water)	70	28	OS	(0.4000,0.6000,0.0000,0.0000)*
99	Father employment(Employee,paid.emp)	370	149	OS	(0.3864,0.5972,0.0135,0.0027)*
25	Kaccha hh (no kaccha hh)	3265	1169	OS	(0.3249,0.6419,0.0290,0.0039)*
13	Mother education (yes)	2333	208	OS	(0.0792,0.9108,0.0051,0.0047)*
7	Gender (Male)	8590	1295	OS	(0.0987,0.8492,0.0353,0.0166)*

*denoted terminal nodes or decision node

Classification tree gives as the information about internal nodes which are shown in the above table .

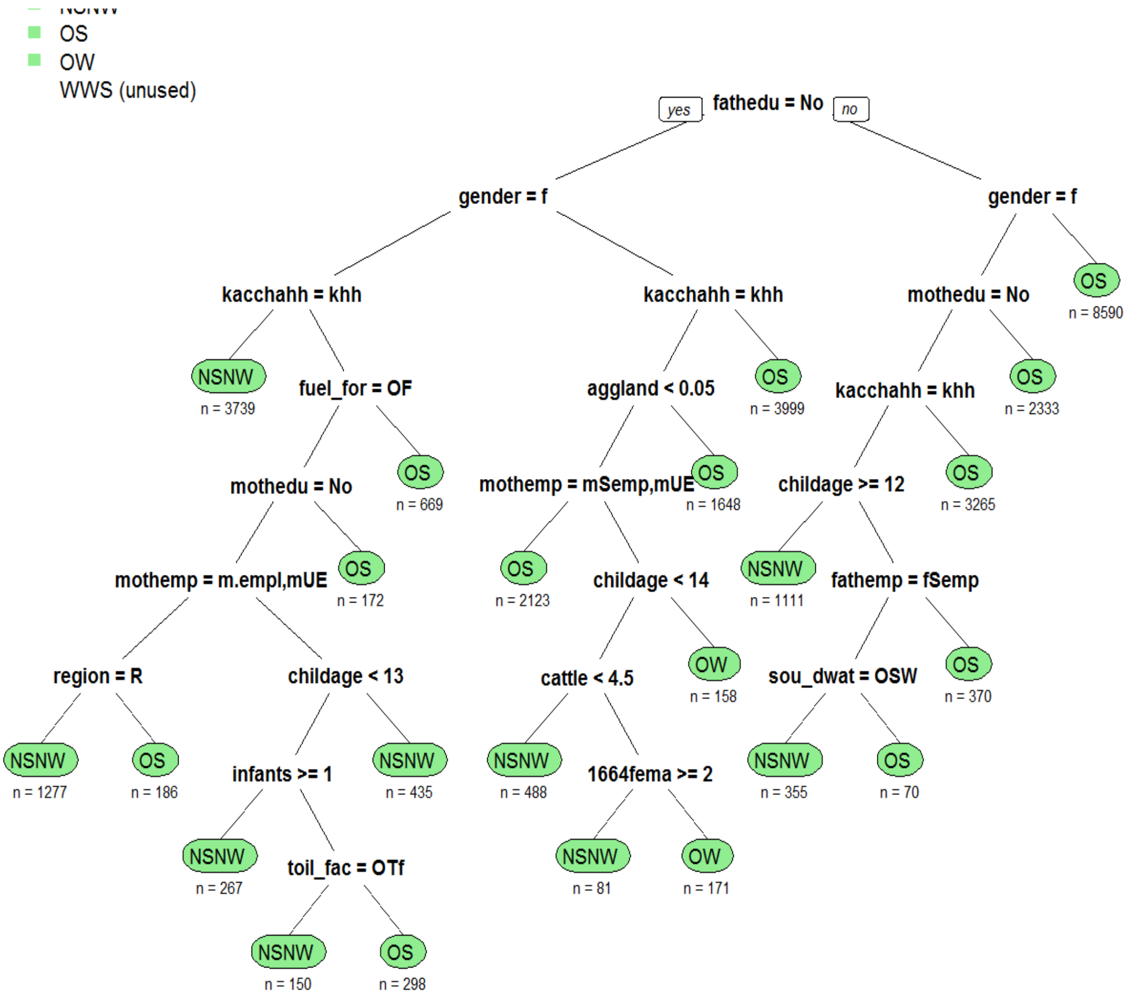
Extra benefit of classification tree method is that it can display the results in the form of tree as like graph and which are more easily interpretable. Tree like graph gives us same explanation as given by the internal nodes of classification tree in the table.

Right branches of tree: most important variable in the model is father education. Child's father is educated and child's gender is male then predicted that child are involve in only school activity. Those child's whose father is educated and child's gender is female but mother is uneducated ,they live in the kaccha house and child's age is greater than equals to 12 years then predicted that child are involved in no schooling no work activity . Those child's having less than 12 year old ,father is self -employee but in the house source of drinking water is maybe tube well, pumped water etc. then predicted that child are involved in no schooling no work activity. The child's who is less than 12 years old ,father is self -employee but in the house source of drinking water is piped water then predicted that child is involved in only school activity.

Middle branches of tree: father is non-educated and child's gender is male and they live in the pakka house then its predict that child are involve in only school activity. father is non-educated and child's gender is male ,they live in the kaccha house but they can own less than 0.05 hectares agriculture land and child's mother are also self -employee or unemployment then predicted that child are involve in only school activity and if father is non-educated and child's gender is male ,they live in the kaccha house and they can own less than 0.05 hectares agriculture land but child's mother is paid employee ,unpaid worker etc and child age having no less than 14 then predicted that child are involve in only work activity. father is non-educated and child's gender is male ,they live in the kaccha house and they can own less than 0.05 hectares agriculture land but child's mother is paid employee ,unpaid worker etc and child age having less than 14,they also own less than 4.5 number of cattle's then this information predicts that child are involved in no schooling no work activity. If father is uneducated and child’s gender is male ,they live in the kaccha house and they own less than 0.05 hectares agriculture land but child’s mother is paid employee ,unpaid worker etc and child age having less than 14,they also own greater than 4.5 number of cattle’s and number of female between the age group 16-64 is not greater and equal to 2 then predicted that child is involved in only work activity.

Right branches of the tree: father is non-educated and child’s gender is female they live in the kaccha house then predicted that child is involved in no schooling no work activity and with the same scenario but those household who lives in the pakka house and uses gas as the fuel for cooking then predicted that child is involved in the only schooling activity. Father is uneducated and child’s gender is female they live in the pakka house ,they used other fuel for cooking like wood, coal and child’s mother is also non -educated but they are employee in any institution or they can do nothing(unemployed),those household belong to rural area then predicted that child is involved in no schooling no work activity and those household belongs to urban region then predicted that child is involved in the only schooling activity. father is non-educated and child’s gender is female they live in the pakka house ,they use other fuel for cooking like wood, coal and child’s mother is also non -educated but they are employee ,paid worker and unpaid employee in any institution and child’s age is not less than 13 years then classify that child belongs to no schooling no work activity and if child’s age is less than 13 years with number of infant in family is greater equal to one then model classify that child belongs to no schooling no work activity. If pervious scenario is same but number of infants in family are not greater than equal to one and in house they have proper sanitation system then predicted that child is involved in the only schooling activity and if there is no proper sanitation system in house then predicted that child is involved in no schooling no work activity.

FIGURE 4. 8: CLASSIFICATION TREE FOR 10-14 YEARS CHILD ACTIVITIES



4.7.1 Confusion Matrix of CT (10-14) by Using Training Dataset

Table represents the prediction with the help of confusion matrix by using the training dataset. The diagonal have correct observations of prediction and off diagonal have misclassify observation. It has been observed that 15.19 %predicted response variable belongs from NSNW activity , 54.41% of predicted response variables belong from OS activity ,0.475% of predicted response variables belong from OW activity and only 0% predicted response belong from WWS activity whereas ,total accuracy of the model by using training dataset is 70.09%, it means that the overall error of model is 29.91%.

TABLE 4. 16: CONFUSION MATRIX OF CLASSIFICATION TREE FOR TRAINING DATASET

predicted	NSNW	OS	OW	WWS
NSNW	4857	2252	732	62
OS	4637	17388	1328	370
OW	72	88	152	17
WWS	0	0	0	0

4.7.2 Prediction by Using Testing Dataset

The overall error for testing dataset model is 29.56%. therefore, the model is fit good on both training and testing dataset and there is no problem of overfitting

4.8 Random Forest of Child Age Between 10-14

Further the accuracy (prediction) of classification tree model is improved by using ensemble methods like random forest.

4.8.1 Results of Random Forest

In the figure 5.9 represent the error of random forest ,Result of random forest give us the separate error rate for each classes. Similarly ,the plot of random forest also shows the error for different classes and the out of bag sample(OOB) over the number of trees. Out of bag sample is the overall error rate of the model ,out of bag sample is in black color and other are in different color such that error of NSNW activity is in red color, error of OS in green color ,error of OW is in blue color and error rate of WWS is in light blue color.

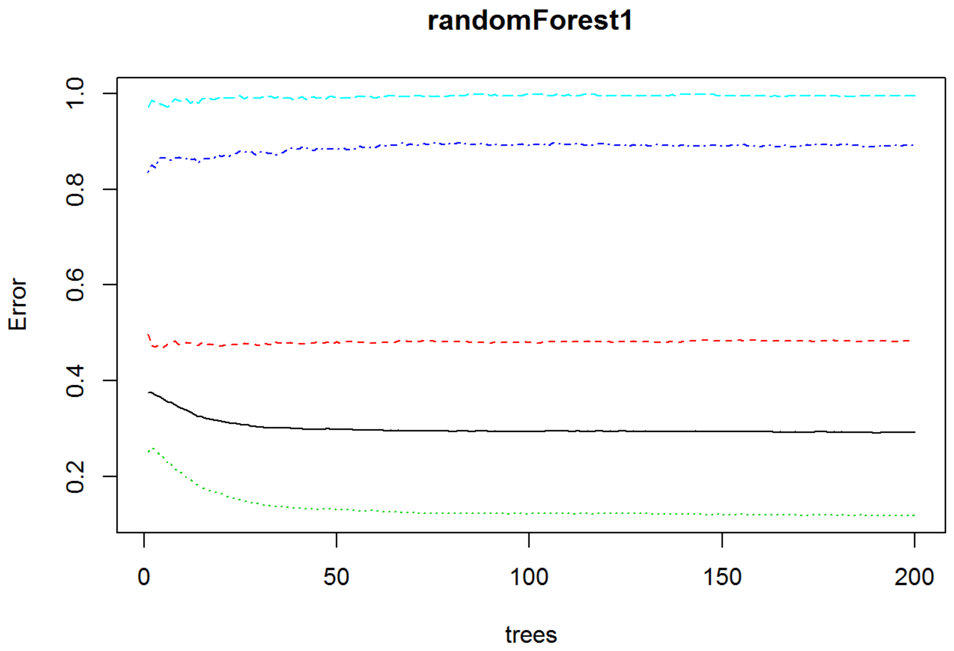


FIGURE 4. 9: ERROR OF RANDOM FOREST FOR DIFFERENT CLASSES USING TRAINING DATA

In this method number of trees are choose randomly, i-e, 200 .As the number of trees increases there is no effect on the error rate.it can be seen from figure5.9 out of bag line is stable and smooth our 20 numbers of tree and the error rate are equals to the 29.28%.

Confusion matrix of random forest is provided in tables 5.15. we have observed that 15.207% correct prediction that are belongs from the NSNW group ,54.748% correct prediction that are belongs from the OS group , 0.7599% correct prediction that belongs from the OW group and it has been observed that only 0.0062% correct prediction come from the WWS groups. Error of group 1(NSNW) is 48.30%, group 2(OS) is 11.94% , group 3(OW) is 89.08 % and group 4(WWS) 99.56%. The list of important variable obtain by random forest are represented in table 5.15 below.

TABLE 4. 17 :SUMMARY OF RANDOM FOREST MODEL

<table>
<tr><td colspan="6"><b>Types of Random forest : Classification</b></td></tr>
<tr><td colspan="6"><b>Number of Trees = 200</b></td></tr>
<tr><td colspan="6"><b>No. of variable tried at each split: 4</b></td></tr>
<tr><td colspan="6"><b>OOB estimate of error rate : 29.28%</b></td></tr>
<tr><td colspan="6"><b>Confusion Matrix :</b></td></tr>
<tr><td></td><td><b>NSNW</b></td><td><b>OS</b></td><td><b>OW</b></td><td><b>WWS</b></td><td><b>Class. Error</b></td></tr>
<tr><td><b>NSNW</b></td><td><b>4863</b></td><td><b>4407</b></td><td><b>138</b></td><td><b>0</b></td><td><b>0.4830995</b></td></tr>
<tr><td><b>OS</b></td><td><b>2255</b></td><td><b>17507</b></td><td><b>118</b></td><td><b>2</b></td><td><b>0.1194548</b></td></tr>
<tr><td><b>OW</b></td><td><b>768</b></td><td><b>1214</b></td><td><b>243</b></td><td><b>1</b></td><td><b>0.8908356</b></td></tr>
<tr><td><b>WWS</b></td><td><b>51</b></td><td><b>373</b></td><td><b>35</b></td><td><b>2</b></td><td><b>0.9956616</b></td></tr>
</table>

However ,we are more interested to know which variable is more important in the building of child activity model .two measures, (i) mean decrease accuracy and (ii)mean decrease Gini are commonly used in the investigation of important variables under random forest. Figure 5.10 is represent the plot of important variables.

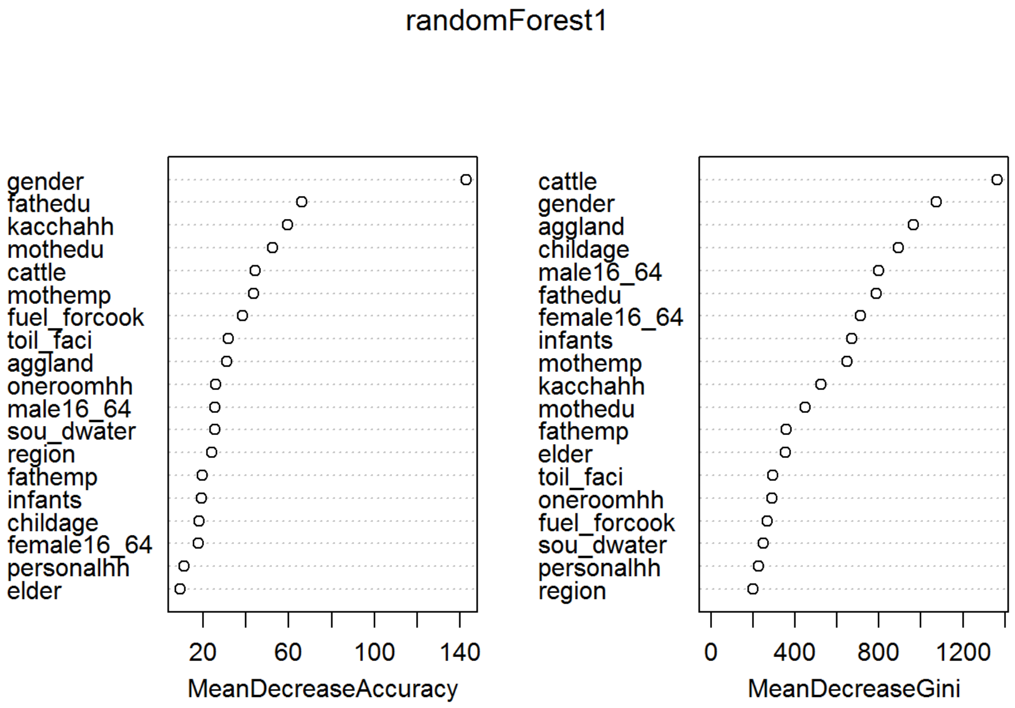


FIGURE 4. 10 : PLOT OF IMPORTANT VARIABLE

4.8.2 Confusion Matrix of RF(10-14) by Using the Testing Data

In table 4.16, confusion matrix is represented for test dataset. It is noted that 24.95 % correct prediction that belongs from the NSNW group , 60.09% correct prediction that are belongs from the OS group ,4.91% correct prediction that belongs from the OW group and we can observed that only 0.61% correct prediction come from the WWS groups. Total accuracy of model is 90.58% it means that the overall error of that model is 9.42%. by using testing dataset error of random forest is less as compare to the classification tree model. Therefore ,random forest has improved the accuracy

TABLE 4. 18:CONFUSION MATRIX OF RF(10-14)

predicted	**NSNW**	**OS**	**OW**	**WWS**
NSNW	**5316**	**413**	**139**	**19**
OS	**979**	**12801**	**306**	**124**
OW	**8**	**9**	**1048**	**9**
WWS	**0**	**0**	**0**	**132**

4.9 Linear Discriminate Analysis

LDA results gives us prior probability ,group means and coefficients of LDA . π_1=0.294790 , $\pi_2 = 0.62255$, π_3=0.06856, π_4=0.014084 are prior probability of child activities which are categorize in four groups .these groups are no schooling no work(NSNW) ,only schooling(OS),only work (OW) and work with schooling(WWS) ; in other words , 29.47% of the training observation goes to no schooling no work activity ,62.25 % belongs to school activity and 6.85% relates to work activity and 1.48% observation are belongs to the work with schooling activity (WWS).

This technique can also deliver the group mean of all categories. group means are actually average of each predictor with in the each class. Coefficient of linear discriminats output delivers the linear combination of all variables that are used for the LDA decision rule . if $+0.0496 * regionU + 1.02822 * sexM - 0.22508 * age - 0.108426 * infants + 0.036083 * elder + 0.06088041 * male16 - 64 + 0.09148 * female16 - 64 - 0.09263 * Fatherpaidemp - 0.23709 * Fatherselfemp - 0.4360 * fatherunpaidwork + 0.31454 * motherpaidemp + 0.68996 * motherselfemp + 0.64816 * motherunemp +$

$0.27589 * motherunpaidwork + 0.35878 * pipedw - 0.3472 * OTF - 0.30287 * OFU + 0.6678 * NKH - 0.30135 * oneroomyes + 0.14039 * personalhhyes - 0.00506 * cattle + 0.000721 * aggland + 0.6753 * motheredu + 0.88177 * fatheredu$ than lda is classifier will predicted that they are belongs to no work no schooling activity and if it is less then lda classifier will predict that child activities are not belongs to no work no schooling activity its belongs to work with schooling activity. similarly ,if $+0.22871 * regionU + 1.26085 * sexM + 0.274301 * age + 0.08132 * infants + 0.0948 * elder - 0.12520 * male16 - 64 - 0.066 * female16 - 64 - 0.1118 * Fatherpaidemp + 0.2079 * Fatherselfemp + 0.3090 * fatherunpaidwork - 1.5372 * motherpaidemp - 1.5128 * motherselfemp - 2.3535 * motherunemp - 0.6828 * motherunpaidwork - 0.1790 * pipedw + 0.1545 * OTF - 0.1553 * OFU + 0.0349 * NKH - 0.0750 * oneroomyes - 0.1839 * personalhhyes + 0.0115 * cattle - 0.0008 * aggland + 0.1803 * mothereduyes + 0.07215 * fathereduyes$ is larger than LDA classifier will predicted that child's activities are belongs to only schooling activity and if it is less than they are belongs to work with schooling activity. if $+0.2340 * regionU + 0.6033 * sexM - 0.0989 * age + 0.2222 * infants - 0.05325 * elder + 0.1529 * male16 - 64 - 0.01777 * female16 - 64 + 0.6740 * Fatherpaidemp + 0.8433 * Fatherselfemp + 5.9143 * fatherunpaidwork + 7.3376 * motherpaidemp + 9.8312 * motherselfemp + 9.093 * motherunemp + 9.0936 * motherunpaidwork + 0.0781 * pipedw - 1.1210 * OTF + 0.7443 * OFU - 0.5805 * NKH - 0.1846 * oneroomyes + 02701 * personalhhyes - 0.0166 * cattle + 0.0929 * aggland - 0.5611 * motheredu - 0.04699 * fathereduyes$ is larger than LDA classify that child's activities are belongs to the only work activity and if it is less than activities are not belongs to the only work activity they are belong (i.e. no schooling no work activity).

Linear discriminant plots are obtained through LDA functions. In these plots show that how LDA classify four groups along with the histogram. These plots also show that the separation of four groups along with the overlying areas ,these overlying or error are possible when they predicting the groups

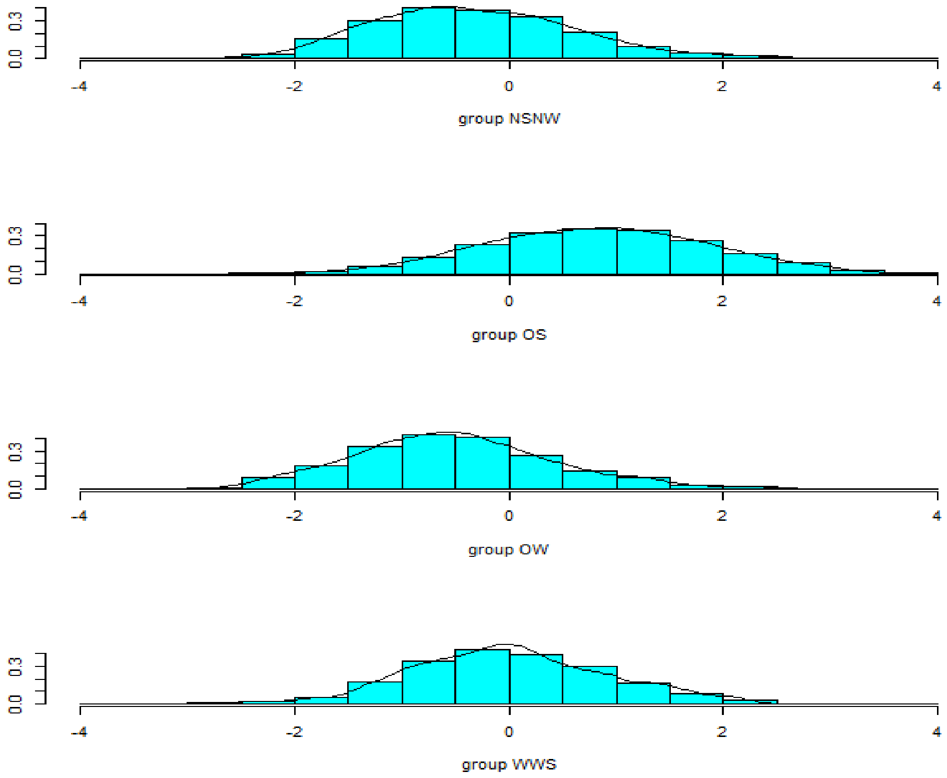


FIGURE 4. 11:PLOT OF LDA FUNCTION

4.10.Confusion Matrix of LDA(10-14) by Using the Training Dataset

Table 5.17, represents the prediction from confusion matrix over the training dataset. It has been observed from the table that 14.46% predicted response variable belongs from NSNW activity ,53.82% of predicted response variables belong from OS activity ,1.21% of predicted response variables belong from OW activity and only 1 predicted response belong from WWS activity whereas ,total accuracy of LDA model by using training dataset is 69.51% it means that the overall error of LDA model is 30.49%

TABLE 4. 19: CONFUSION MATRIX OF LDA FOR TRAINING DATASET

predicted	NSNW	OS	OW	WWS
NSNW	4601	2287	725	54
OS	4490	17120	1068	328
OW	286	395	387	65
WWS	0	1	1	1

4.10 Evaluating the Above Techniques on The Model of Child Activities Between the Age 10-14

In this section we are evaluating or comparing the three techniques for testing/unseen dataset that are multinomial logistic regression(MLR),Classification and regression tree (CART) and Linear discriminant analysis (LDA) on the basis of two methods that is confusion matrix and ROC analysis both tell us the accuracy of the models .

4.10.1 Accuracy Based on Confusion Matrix for all Techniques

TABLE 4. 20; ACCURACY RATES OF CONFUSION MATRIX FOR ALL TECHNIQUES

Techniques	Testing dataset
MLR	69%
CT	70%
LDA	69%

4.10.2 Accuracy Measurement Based on ROC Analysis

The Roc analysis of MLR, CT and LDA are presented in 5.12 ,5.13 and 5.14.The middle(black) line show that the area under the curve and other lines shows that specificity and sensitivity of each group of activity .area under the curve is actual measure of accuracy of this model. The AUC of this model is 0.6023.

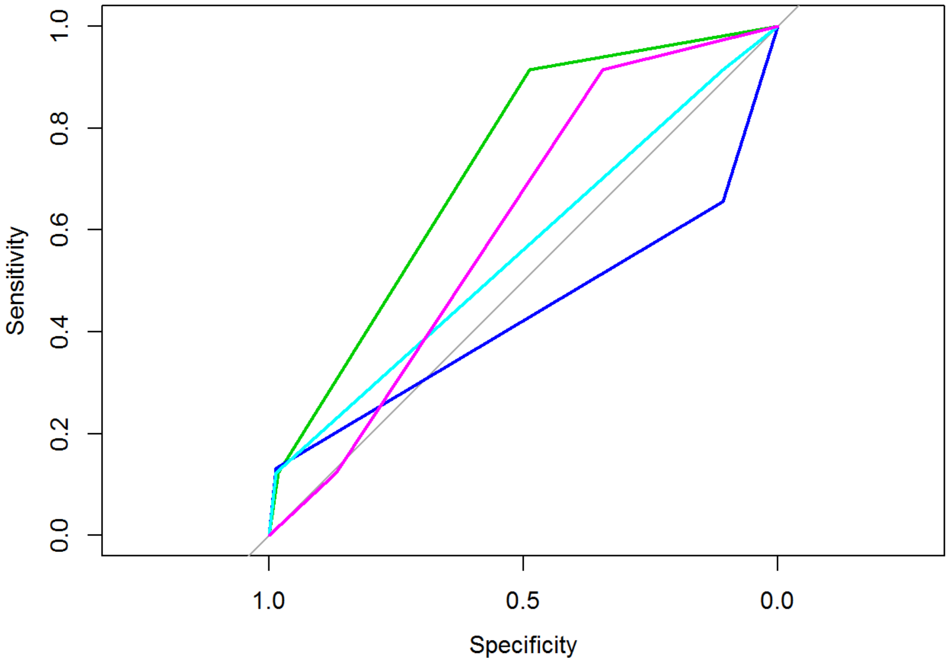


FIGURE 4. 12: MULTI-ROC ANALYSIS FOR MLR MODEL

The middle(black) line show that the area under the curve and other lines shows that specificity and sensitivity of each group of activity .area under the curve is actual measure of accuracy of this model .the AUC of this model is 0.5956/0.60

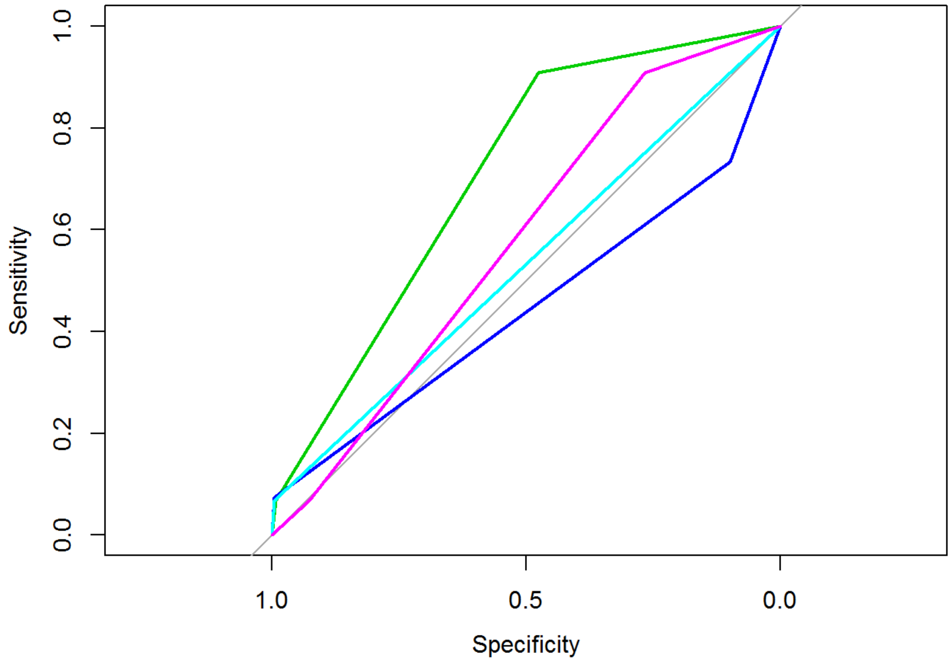


FIGURE 4. 13: MULTI-ROC ANALYSIS FOR CT MODEL

The middle(black) line show that the area under the curve and other lines shows that specificity and sensitivity of each group of activity .area under the curve is actual measure of accuracy of this model. The AUC of this model is 0.6042.

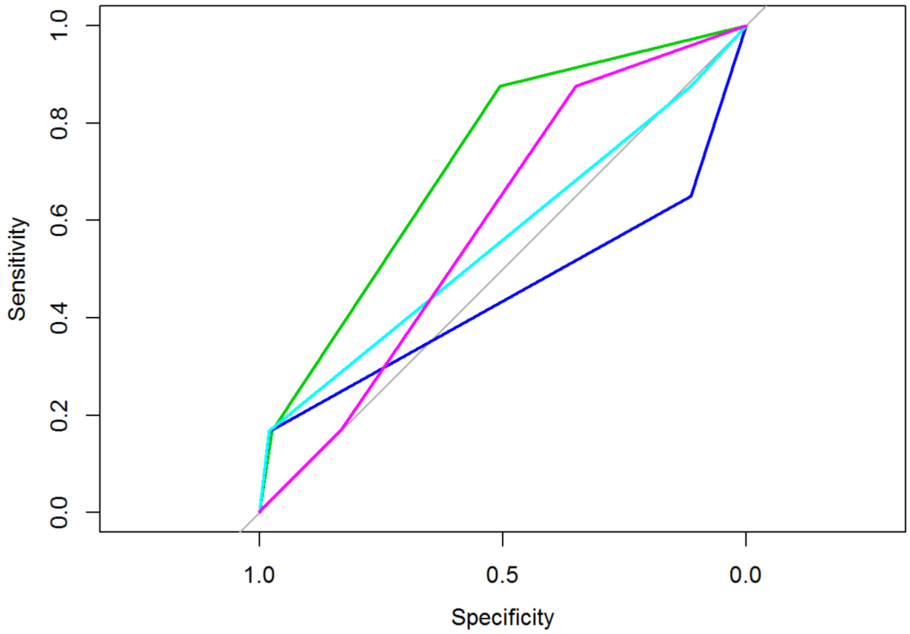


FIGURE 4. 14: MULTI-ROC ANALYSIS FOR LDA (10-14)

Our results indicated that when we compare the three techniques for child activities between the age group 10-14 ,conclude that the machine learning technique (classification tree method) is better than multinomial logistic regression and linear discriminate analysis on the basis of both method.

CHAPTER 5

CONCLUSION AND SUMMARY

In this study we have compared two approaches that is machine learning and classical approach for classify the child activities. In machine learning approach the methods of classification are classification tree and linear discriminate analysis. In classical approach the method of classification is multinomial logistic regression .we also investigated the accuracy of these models in terms of prediction come from confusion matrix and also examined the area under the curve in the Roc analysis.

We had two groups of children which were divided according to age groups .in the model of child activities between the age group of 4-9 year had three categories and in the model of child activities between the age group of 10-14 year had four categories to examine the performance of MLR , LDA and CT.

We concluded that ,when compare the performances of MLR,CT and LDA models in term of accuracy on the child activities between the age group of 4-9 then both Multinomial logistic regression and linear discriminant analysis give similar results but classification tree model performed better than MLR and LDA. The overall accuracy of classification tree model was good ,accuracy come from confusion matrix and area under the curve of classification tree model was greater than the MLR and LDA. LDA could perform better than MLR and CT but this was effected due to the violation of the Gaussian distribution (multivariate normality) assumption for child activities dataset and also effected due to the heterogenous covariances matrices

Also ,when we comparing the performances of MLR ,CT and LDA model on the child activities between age group of 10-14 then we concluded that the overall accuracy of classification tree is slight greater than the MLR and CT, accuracy of confusion matrix indicated that there is difference in the performances of three model ,so classification tree is performed better than MLR and LDA in the child activities between the age group of 10-14, but when we see the accuracy in to the account of area under the curve of ROC analysis indicates that the accuracy of three model were almost similar so there is no significant difference between the performance of three model.

At the end , we conclude that the classification tree is better technique and helpful for giving the information about which variable is more important and this information can easily see in the tree diagram .according to the classification tree, age , father education ,gender ,kaccaha house, mother education ,one room house and toilet facility is important variable in the model of child activities between the age group of 4-9 and Father education ,child gender, kaccha house, fuel for cooking mother education ,mother employment ,region ,child'sage, infants, toilet facility, aggland, cattle,16-64female,source of drinking water, father employment are important variables in the second model of the child activities that is between the age of 10-14.

5.1. Policy Recommendation

1. In the light of our empirical evidences , child gender in both model have significant effect on the decision of child activities. Many developing countries like Pakistan ,facing gender discrimination problem. Our society is also male oriented and cannot give the preference to girls' education. As we also know that in Pakistan more than half of population is consisting on the woman, so without woman contribution in economic activities country cannot move towards developed stage. Therefore, it is suggested that gender disparity should be minimized through public awareness about girls' education. Secondly we should make separate setup for girls schooling particularly in that area where people cannot sent their girls child just because of combine schooling. Thirdly

, we should provide the free vocational training to the girls with the basic education. Free vocational training in school for girls can beneficial in multi way. In the free vocational training girls can learn lot of skill’s. On the basis of skills girls can pursue own business at home level afterward they extend at high level. They can earn from business and support their family as well as indirectly they can contributing in economic level. Due to education from vocational skill parents can also satisfied for girls future and sent their girls child in to schooling because they know that with getting basic education girl child can also learn a lot of skills which are financially beneficial for family.

2. Woman education in both model have significant effect on the decision of child activities. We have to focus on girl’s education because in future girls can play important role as woman. And every educated women knows the worth of education or they engage their child in school going activity. It has an increasing effect on human capital through the education.

REFERENCES

Ali, K., & Khan, R. E. A. (2004). Simultaneous decision making of child schooling and child labour in Pakistani urban households.

Balogun, O. S., Akingbade, T. J., & Oguntunde, P. E. (2015). An assessment of the performance of discriminant analysis and the logistic regression methods in classification of mode of delivery of an expectant mother. *Mathematical Theory and Modeling*, *5*(5).

Burki, A. A., Fasih, T., & Din, M. U. (1998). Households' Non-leisure Time Allocation for Children and Determinants of Child Labour in Punjab, Pakistan [with Comments]. *The Pakistan Development Review*, 899-914.

Cockburn, J. (2001). Child labour versus education: poverty constraints or income opportunities?. *Center for the Study of African Economies, Oxford University*.

Dar, A., Blunch, N. H., Kim, B., & Sasaki, M. (2002). Participation of children in schooling and labor activities: A review of empirical studies. *World Bank, Social Protection Discussion Paper*, *221*.

Duraisamy, M. (2000). Child schooling and child work in India. *Department of Humanities and Social Sciences, Indian Institute of Technology, Madras (Processed)*.

Heltberg, R., & Johannesen, N. (2002). *How Parental Education Affects Child Human Capital: Evidence from Mozambique, Institute of Economics, University of Copenhagen*. Discussion paper 02-04.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons. (Book)

Iram, N., Hussain, Z., Anwar, S., Hussain, I., & Akram, W. (2008). Determinants of child school choice in Punjab: Policy implications. *European Journal of Scientific Research*, *23*(2), 285-293.

Jensen, P., & Nielsen, H. S. (1997). Child labour or school attendance? Evidence from Zambia. *Journal of population economics*, *10*(4), 407-424.

Khan, R. E. A. (2003). Children in different activities: child schooling and child labour. *The Pakistan Development Review*, 137-160.

Khan, R. E. A., Khan, T., & Sattar, R. (2010). A comparative analysis of rural and urban child labor in Pakistan.

Khanam, R. (2005, March). Impact of child labour on school attendance and school attainment: Evidence from Bangladesh. In *Population Association of America Annual Meeting (PAA 2005)* (pp. 1-40). University of Southern Queensland.

Montgomery, M. E., White, M. E., & Martin, S. W. (1987). A comparison of discriminant analysis and logistic regression for the prediction of coliform mastitis in dairy cows. *Canadian journal of veterinary research*, *51*(4), 495.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, *50*(3), 559-569.

Nkamleu, G. B., & Kielland, A. (2006). Modeling farmers' decisions on child labor and schooling in the cocoa sector: a multinomial logit analysis in Côte d'Ivoire. *Agricultural Economics*, *35*(3), 319-333.

Okurut, F. N., & Yinusa, D. O. (2009). Determinants of child labour and schooling in Botswana: evidence from 2005/2006 labour force survey. *Botswana Journal of Economics*, *6*(10), 15-33.

Parikh, A., & Sadoulet, E. (2005). The effect of parents' occupation on child labor and school attendance in Brazil.

Pohar, M., Blas, M., & Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: a simulation study. *Metodoloski zvezki*, *1*(1), 143.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, *73*(364), 699-705.

Sathar, Z. A., & Lloyd, C. B. (1994). Who gets primary schooling in Pakistan: Inequalities among and within families. *The Pakistan Development Review*, 103-134.

Tapak, L., Mahjub, H., Hamidi, O., & Poorolajal, J. (2013). Real-data comparison of data mining methods in prediction of diabetes in Iran. *Healthcare informatics research*, *19*(3), 177-185.

Worth, A. P., & Cronin, M. T. (2003). The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *Journal of Molecular Structure: THEOCHEM*, *622*(1-2), 97-111.

Yamada, K. (2011). Family background and economic outcomes in Japan.