KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUMASI, GHANA

Classification trees as a tool for decomposing inequalities in under - five mortality in Ghana

By

YEBOAH DANIEL (B.Sc.)

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF M.PHIL MATHEMATICAL STATISTICS

May 28, 2014

Declaration

I hereby declare that this submission is my own work towards the award of the M. Phil degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgement had been made in the text.

Yeboah Daniel (PG8066912)	.....	.....
Student	Signature	Date

Certified by:

Nana Kena Frempong	.....	.....
Supervisor	Signature	Date

Certified by:

Prof. S. K. Amponsah	.....	.....
Head of Department	Signature	Date

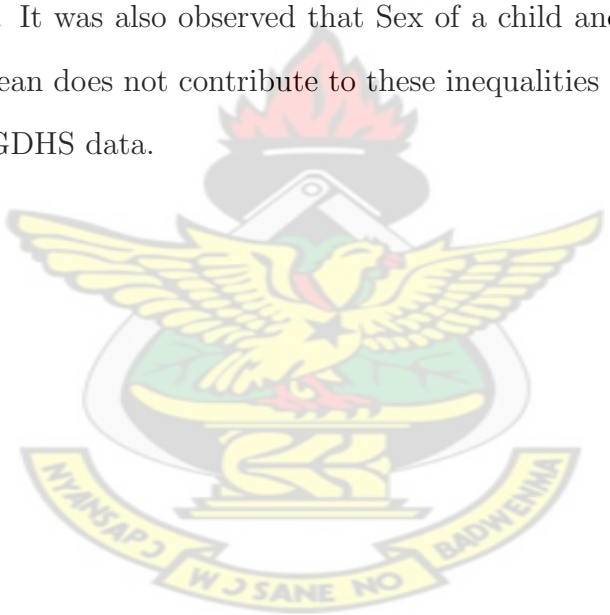
Dedication

This work is dedicated to my parents for their care and support

Abstract

The study explores the socio-economic and demographic inequalities that exist in under-five mortality using Ghana DHS 2008 data.These inequalities were investigated using Classification trees developed by Breiman et al (CART). The data was analyzed using the Rpart of the R software 3.01

At the end of the analysis, total number of children ever born by mothers, preceding birth interval, region of residence, ethnicity, highest educational level and wealth index were identified to contribute positively to these inequalities in under-five mortality. It was also observed that Sex of a child and whether a child was born by cesarean does not contribute to these inequalities in the under-five mortality in the GDHS data.



Acknowledgement

I am most grateful to the Almighty God for His guidance in the realization of this thesis. I also wish to express my deepest gratitude and appreciation to my supervisor, Nana Kena Frempong for his mentorship and tutelage throughout the period.

It is my pleasure to thank Mr. Emmanuel Nakuah, lecturer at the School of Medical Sciences, KNUST and Dr. Peter Twumasi, lecturer at the Department of Biochemistry, KNUST for their support during the research period.

I would like to appreciate the support of my colleagues; Opoku Gyabaah, Daniel Lipi and Mr. Peter Kwasi Sarpong for their friendship, motivation, moral support and assistance.

Finally,My deepest thanks and love go to my parents Deacon Kingsley Sarfo and Janet Osaah whose love,encouragement and trust in me has always given me strength and confidence throughout my life.

Contents

List of Abbreviations

UNDP	United Nations Development Program
DHS	Demographic and Health Survey
WHO	World Health Organization
GHS	Ghana Health Service
CSDH	Commission on Social Determinants of Health
GNI	Gross National Income
GSS	Ghana Statistical Service
MDG	Millennium Development Goal
SRH	Self - Rated Health
PHC	Primary Health Care
QLI	Quality of Life Index
UNICEF	United Nations International Children's Emergency Fund
UN	United Nations
USAID	United States Agency for International Development
CART	Classification and Regression Tree
ICF	Inner City Fund
USAID	United States Agency for International Development
UNFPA	United Nations Population Fund
DANIDA	Danish Development Agency

List of Tables

Chapter 1

Introduction

1.1 Background to the Study

Health is certainly a basic need in all human societies. Equity which is defined as "the absence of potentially remediable, systematic differences in one or more aspects of health across socially, economically, demographically or geographically defined population, groups or subgroups" (International Society for Inequity in Health,2010) is a very important aspect of health. This implies that, health equity is a well accepted ethical and human rights principle; that is all human beings are entitled to have the highest attainable level of health care.

International Organizations such as the World Health Organization (WHO) and United Nations Development Program (UNDP) have viewed health as the most important goal for human development and the fundamental indicator of social development (Feng and Yangyang, 2000). Health is not only instrumental in enabling people to earn a living and to enjoy the fruits of their labor, but is an important element of well-being in its own right. It should therefore be accepted that the health status of a nation is an important indicator of the well-being of its citizenry. According to the WHO report ((UN/MDG, 2012)), the objectives of good health are in two folds. "The best attainable average level (goodness) and the smallest feasible differences among individuals and groups (fairness)." An effort to improve health in developing countries faces many challenges. One of the main challenges is under-five mortality.

Under-five mortality rates are basic indicators of a country's socio-economic sit-

uation and quality of life, as well as specific measures of health status. Each year in the global village, approximately four million infants die within the first four weeks of their birth. This number is conservative estimate given that children of color in developing countries account for 98% of neonatal deaths worldwide and two-thirds of the world's population, many under-five deaths go unseen. Experts estimates that these four million neonatal deaths comprise thirty six percent of the mortality rate for all children younger than five years of age. In the year 2000, 99 percent of the 10.9 million childhood deaths occurred in the developing countries (WHO, 2001). One of the Millennium Development Goals is to reduce by two-thirds, infant and under-five mortality rate by the year 2015 ((NDPC/GOG, 2012)).

Despite the fact that several studies has been done on under - five mortality and its numerous determining socio-economic factors, they have not reconciled a common solution about how this problem needs to be solved in developing countries of which Ghana is an emerging economy. The adoption of uniformly agreed development policies for the reduction of high rates of child mortality would not be a wise decision in many developing countries given the different socio-economic, political, environmental and cultural settings.

On the other hand, the findings of some research studies have shown that certain policies pursued by some developing countries have not been effective in battling high rates of child mortality. It goes without saying that policies for reducing high rates of child mortality in developing countries must take into consideration economic, social, cultural, environmental and political factors. However, this does not mean that there should not be any attempt to create a universal theoretical framework that could explain which factors are the most important and which policy actions could be the most effective in decreasing under-five mortality at certain stages of economic development. This framework would help to shape

different policies to achieve a fast and sustain decrease in child mortality in developing populations.

So measuring health inequality in child mortality is the main part of assessing the performance of a health system in a country. Despite an improvement in many health indices in different countries during the past decades, health inequality has not only remained but also increased in some of them. Investigating the inequality of a binary dependent variable (example health) has received attention mainly in the last decade. One way to think of this inequality is in terms of a univariate distribution. Some individuals have high health, others have low health and health inequality is intended to indicate the extent of dispersion of "health" within the population.

In this study, the focus is on this univariate approach using a fatal health outcome, child mortality resulting in a binary index with "no" denoting a surviving child and "yes" denoting a child that died within its first five years of birth. It can be noticed that the WHO defined inequality as "differences in health status which are unnecessary and avoidable, but in addition are considered unfair and unjust" (Whitehead and Dahlgren (1992)). For example, it was estimated in the GDHS data (2008) that mortality levels in the rural areas are consistently higher than those in urban areas. In the ten-year period before the survey of GDHS, infant mortality in rural areas was 56 deaths per 1,000 live births, compared with 46 deaths per 1,000 live births in urban areas. Also differences in mortality by region are marked. The infant mortality rate varies from 36 deaths per 1,000 live births in Greater Accra to 97 deaths per 1,000 live births in the Upper West region. Expectant mothers' education is inversely related to a child's risk of dying. Children of women with no education (61 deaths per 1,000 live births) are much more likely to die in the first year than children of women with middle/ JSS education (46 deaths per 1,000 live births). These might be considered as

inequalities. For this author, health equity implies that ideally everyone should have a fair opportunity to attain their full health potential and more pragmatically, no one should be disadvantaged from achieving this potential, if it can be avoided. (Whitehead and Dahlgren, 1992) identifies seven possible determinants of health inequalities:

i natural, biological variation; example, in the GDHS data, it was discovered that under-five mortality rate for male and female children are 93 and 76 deaths per 1,000 live births, respectively. This excess mortality among male children is most likely due to their higher biological risk during the first month of life. Findings from the World Fertility Survey and DHS survey indicate that births to young mothers (under age 20 years) and older mothers (35 years and above) are at an elevated risk of dying.

ii health-damaging behavioral choices;

iii temporary health advantages occurring to one group when it adopts health promoting behaviors (assuming other groups have equal chance of adopting the behavior);

iv health-damaging behaviors where the degree of choice is severely restricted;

v exposure to unhealthy, stressful living and working conditions;

vi inadequate access to essential health and other basic services; and

vii natural selection or health-related social mobility involving the tendency for sick people to move down the social scale.

She does not consider the first three of these as being unjust, while the last four are both avoidable and unjust. Based on this explicitly pragmatic point, she defines equity using two antonyms, 'inequality' and 'inequity'. 'Inequality' refers to systematic, unavoidable, and meaningful differences among members of a population; 'inequity' refers to the existence of variations which are not only

unnecessary and avoidable, but also unjust. Whitehead herself points out that equity does not mean that everyone should enjoy the same level of health and consume services and resources to the same degree. Rather the needs of each individual should be addressed. To describe a situation as inequitable or unjust, it needs to be examined and judged in a larger social context. To summarize, any inequity is an inequality but not every inequality is an inequity. An inequity is an unjust and potentially avoidable inequality.

Having measured inequalities in the health sector, a natural next step is to seek to explain them. Variations in health as an example may be explained by a large number of (unfair) factors like place of residence, religion, occupation, gender, age differences, ethnicity, variation in education, socio-economic status (income), to mention a few. Decomposition allows measuring the contributions of each factor to such inequality. That is how much of the variability is explained by each of the factors.

1.2 Statement of the Problem

Globally, under-five mortality rates has reduced from 88 deaths per 1,000 live-births in 1990 to 57 per 1,000 in 2010. Despite this remarkable progress, disparity in under-five mortality rates still remains high. Children from rural, poorer households and children of less-educated mothers are more likely to die before their first or fifth birthday than children from urban, wealth quintile and well educated mothers. Dealing with these health inequalities has become a challenge in this country and other developing countries as well. In 2005, the WHO initiated the Commission on Social Determinants of Health (CSDH) and its main work was to understand the social determinants of health, how they operate and how they can be changed to improve health and reduce health inequalities. One of their recommendations was to measure and understand health inequality problems within countries and globally, which will be the platform for actions to be

taken.

Although under-five mortality is higher in rural Ghana, the pattern of mortality is more or less the same. Though, according to the GDHS report in 2008, there seems to be a decline in infant mortality rate in Ghana, much attention has not been given to it by our policy makers. Socio-economic factors such as place of residence, ethnicity, occupation, gender, religion, education and social capital (that are unfair and unjust) still has adverse effect on health inequalities in Ghana. This requires decision makers to act in order to reduce health inequalities in under - five mortality. It is against this backdrop that the researcher wants to explore these socio-economic and demographic inequalities that exist in under-five mortality in Ghana.

1.3 Objectives of the study

The objectives of the study are to use classification trees to identify;

- socio-economic and demographic inequalities in under-five mortality in the rural areas
- socio-economic and demographic inequalities in under-five mortality in the urban areas
- further evidence of the differentials in urban-rural under-five mortality.

1.4 Methodology

Classification tree will be used as a statistical methodology to decompose the inequalities in child mortality. The building of classification tree begins with a node, containing all the subjects and then through a process of yes/ no questions, generates descendant nodes. Beginning with the first node, the idea is to find the best variable. The R software (using the Rpart) checks all possible splitting

variables, as well as all possible values to be used to split the node. In choosing the best splitter, the program seeks to maximize the average "purity" of the two child nodes. Through a process of tree building and pruning of trees, an optimal tree is obtained. We propose the strength of each variable in reducing the impurity or inequality as the measure which defines the variable's contribution.

1.4.1 brief country profile

Ghana, a country of Western Africa, situated on the coast of the Gulf of Guinea, has a population of 24,233,431 of which 11,801,661 (48.7%) are males and 12,421,770 (51.3%) are females The population structure is typical of a developing country with about half of the total population below 15 years of age. Women have an average of 4 children, with average number of children per woman ranging from 3.1 in urban areas to 4.9 in rural areas. The gross national income (GNI) per capita in 2011 was US 1,571 according to the Ghana Statistical service. The population is predominantly of African origin, with the Akan tribe consisting of 44 percent of the population, the Moshi-Dagomba 16 percent, the Ewe 13 percent, the Ga-Adangbe 8 percent, the Yoruba 1.3 percent, and European and other nationalities less than 1 percent (www.nationsencyclopedia.com). It has an area of 239,540 square kilometers (92,486 square miles). Water occupies 8520 square kilometers (3,290 square miles) of the country, primary Lake Volta. Ghana has ten regions: the Northern, Upper West, Upper East, Volta, Ashanti, Brong Ahafo, Eastern, Central, Western and Greater Accra. Also it has 170 districts (9-27 per region), 800 sub-districts and 25,672 communities, its capital is Accra. English is the official language with the other main languages being Akan, Moshi-Dagomba, Ewe, and Ga. The country comprises of Christians, Muslims and Traditionalist.

1.5 Justification of the Work

The study would seek to provide the needed statistical evidence to justify the success or failure of the set target with respect to Ghana Demographic and Health Survey data. On the basis of available empirical evidence, this study would seek to furnish decision makers and other stakeholders with vital information regarding the inequalities in socio-economic factors of child mortality in Ghana for possible policy interventions.

Additionally, this study would also contribute to knowledge in the use of classification trees in the area of child mortality and other related areas with a view of stimulating other research.

1.6 Limitations of the Study

Like any research work, this study is not without limitations. The Ghana DHS 2008 report outlined only the estimates of the deaths of under-five children without the causes of these deaths. Furthermore, the issue of missing information in data resulted in the dropping of some respondents from the final data thereby decreasing the sample size used for the final analysis.

1.7 Organization of the Thesis

The thesis consists of five chapters. Chapter one gives a general background of the thesis and how it was carried out. It includes a general introduction and inequalities in under-five mortality in Ghana and the world as whole, statement of the problem under study, objectives of the study, methodology, justification and limitations of the study. In chapter two, selected research works that are related to under-five mortality, inequalities in health and classification trees are reviewed. Chapter three discusses the data and an in depth explanation of the

methods used. In chapter four, the data is explored, analyzed and discussed using the R statistical software to get results. The last chapter presents the conclusions and recommendations of the study.

Chapter 2

Literature Review

This chapter discusses the literatures available on decomposition of inequality in general. It also looks at child mortality and summary of abstracts on various literatures with regard to the model being used and the general working title.

According to the Demographic Health Survey of Ghana (2008), child mortality has declined from 61 deaths per 1,000 live births in the period $10-14$ years before the survey to 50 deaths per 1,000 live births in the period $0-4$ years before the survey. Results from the five GDHS conducted in 1988, 1993, 1998, 2003, and 2008 show a remarkable decline in childhood mortality over the past 20 years as shown in figure 2.0.

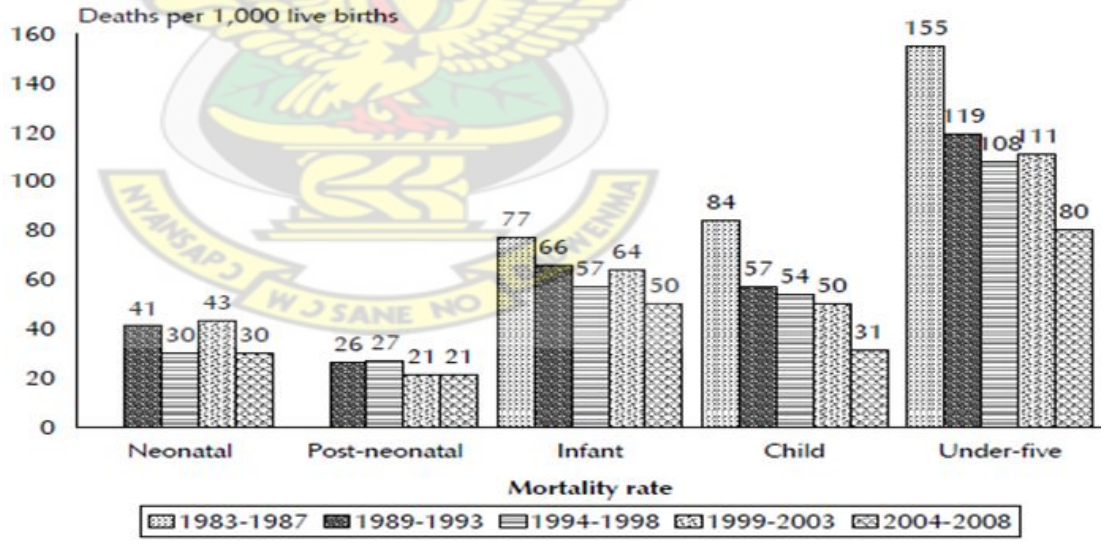


Figure 2.0: mortality trends, Ghana 1988 – 2008

From Figure 2.0, the decline in mortality trends appeared to have halted during the period $1999-2003$ but then declined further during the past five years from 2003 to 2008. This decline in infant and under-five mortality in the five years preceding the 2008 GDHS indicates that the targets set by the Ghana Poverty

Reduction Strategy, an infant mortality rate of 50 per 1,000 and an under-five mortality rate of 95 per 1,000 by 2005 (World Bank, 2003) has been achieved and the MDG's target for childhood mortality is on track.

From the GDHS report, child survival is closely related to socio-economic characteristics of mothers and children. Table 2.0 shows differentials in childhood mortality by four socio-economic variables: Residence, Region, Mother's education, and Household wealth status (quintile).

Table 2.0; Early childhood mortality rate by socio – economic characteristics

Neonatal, post-neonatal, infant, child, and under-five mortality rates for the 10-year period preceding the survey, by background characteristic, Ghana 2008

Background characteristic	Neonatal mortality (NN)	Post-neonatal mortality[1] (PNN)	Infant mortality ($_1q_0$)	Child mortality ($_4q_1$)	Under-five mortality ($_5q_0$)
Residence					
Urban	30	19	49	27	75
Rural	34	23	56	36	90
Region					
Western	40	11	51	(14)	(65)
Central	(47)	(26)	(73)	(38)	(108)
Greater Accra	21	(15)	(36)	(14)	(50)
Volta	26	(11)	(37)	(13)	(50)
Eastern	29	(25)	(53)	(30)	(81)
Ashanti	35	19	54	28	80
Brong Ahafo	27	(10)	(37)	(41)	(76)
Northern	35	35	70	72	137
Upper East	(17)	(30)	(46)	(33)	(78)
Upper West	45	52	97	(50)	(142)
Mother's education					
No education	38	23	61	44	102
Primary	35	20	55	35	88
Middle/JSS	23	23	46	23	68
Secondary+	(38)	(11)	(49)	(15)	(64)
Wealth quintile					
Lowest	31	28	59	47	103
Second	27	18	45	35	79
Middle	44	26	70	34	102
Fourth	31	14	45	25	68
Highest	31	16	46	14	60

Note: Numbers in parentheses are based on 250-499 unweighted exposed persons; an asterisk indicates that a rate is based on fewer than 250 unweighted exposed persons and has been suppressed.

[1] Computed as the difference between the infant and neonatal mortality

According to Table 2.0, differences in mortality by region are marked. The infant mortality rate varies from 36 deaths per 1,000 live births in Greater Accra to 97 deaths per 1,000 live births in the Upper West region. A differential in under-five mortality shows a similar pattern. Mother's education on the other hand, is inversely related to a child's risk of dying. Under-five mortality among children of mothers with no education (102 deaths per 1,000 live births) was substantially higher than under-five mortality among children of women with middle/JSS level

education (68 deaths per 1,000 live births). The direct association between level of education and under-five mortality was also seen in infant mortality. Children in households in the highest wealth quintile have the lowest mortality rates for both child mortality and under-five mortality. Infant mortality was lowest among children in the second, fourth, and fifth wealth quintiles.

The report also revealed that demographic factors are strongly associated with the survival chances of young children. Some of these factors include sex of child, age of mother at birth, birth order, length of preceding birth interval, and size of child at birth. Table 2.10 shows the relationship between childhood mortality and these demographic variables.

Table 2.10: Early childhood mortality rates by demographic characteristics

Neonatal, post-neonatal, infant, child, and under-five mortality rates for the 10-year period preceding the survey, by demographic characteristics, Ghana 2008

Demographic characteristic	Neonatal mortality (NN)	Post-neonatal mortality[1] (PNN)	Infant mortality ($_1q_0$)	Child mortality ($_4q_1$)	Under-five mortality ($_5q_0$)
Sex of child					
Male	36	22	58	38	93
Female	29	20	49	28	76
Mother's age at birth					
<20	43	26	69	43	109
20-29	28	19	46	28	73
30-39	35	25	60	37	95
40-49	(40)	*	*	*	*
Birth order					
1	35	18	52	33	84
2-3	24	24	48	29	75
4-6	36	18	54	38	90
7+	49	31	80	(33)	(110)
Previous birth interval[2]					
<2 years	60	28	88	47	131
2 years	24	30	53	35	86
3 years	30	26	56	29	83
4+ years	23	11	33	25	58
Birth size[3]					
Small/very small	(49)	(35)	(84)	*	*
Average or larger	25	17	42	*	*

Note: Numbers in parentheses are based on 250-499 unweighted exposed persons; an asterisk indicates that a rate is based on fewer than 250 unweighted exposed persons and has been suppressed.

[1] Computed as the difference between the infant and neonatal mortality

[2] Excludes first-order births

[3] Rates are for the five-year period preceding the survey

From Table 2.10, childhood mortality is higher for males than females. The excess mortality among male children was most likely due to their higher biological risk

during the first month of life. Findings from the World Fertility Survey and DHS surveys indicate that births to young mothers (under age 20 years) and older mothers (35 years and over) are at an elevated risk of dying. Results from the 2008 GDHS confirm the expected curvilinear relationship between mother's age at birth and childhood mortality. First births and higher-order birth's typically have an elevated risk of dying. Results from the 2008 GDHS generally confirm this pattern. Mortality among children is negatively associated with the length of the previous birth interval. This was particularly the case when the birth interval is less than two years. The results of the GDHS 2008 indicate that this pattern holds for all levels of childhood mortality except post-neonatal mortality. A child's size at birth has often been found to be an important indicator of the chances of survival during infancy. Majority of births in Ghana take place outside a health facility setting, and these babies are seldom weighed at birth. The mother's assessment of the size of the baby at birth is used as a proxy for birth weight. The GDHS results indicate that among babies assessed by their mother as 'small or very small', infant mortality is twice the level observed for babies assessed as ' average or larger' at birth. The difference in infant mortality between the two groups is largely attributed to neonatal mortality, which is almost twice as high among small or very small babies as among average or larger babies.

(Hosseinpoor et al., 2006) did a study on decomposing socioeconomic inequality in infant mortality in Iran. The objective of their paper was to quantify the determinants' contributions of socioeconomic inequality in infant mortality in Iran. Households' socio-economic status was measured using principal component analysis. The concentration index of infant mortality was used as the measure of socioeconomic inequality and decomposed into its determining factors. They made use of Iranian DHS data for their study. The result shows that, the largest contribution to inequality in infant mortality is as a result of household economic status (36.2%) and mother's education (20.9%). Residency

in rural/urban areas (13.9%), birth interval (13.0%), and hygienic status of toilet (11.9%) also proved important contributors to the measured inequality. It was concluded that, socioeconomic inequality in infant mortality in Iran is determined not only by health system functions but also by factors beyond the scope of health authorities and care delivery system. This implies that in addition to reducing inequalities in wealth and education, investments in water and sanitation infrastructure and programmes (especially in rural areas) are necessary to realize improvements of inequality in infant mortality across society.

(Nkoni, 2011) did a study titled "Explaining household socio-economic related child health inequalities using multiple methods in three diverse settings in South Africa". Their objectives were to study where to measure inequalities in child mortality, HIV transmission and vaccination coverage within a cohort of infants in South Africa. They also applied decomposition technique to identify the factors that contribute to the inequalities in these three child health outcomes. A relative index of household socioeconomic status was developed using principal component analysis. This paper uses the concentration index to summarize inequalities in child mortality, HIV transmission and vaccination coverage. They observed disparities in the availability of infrastructure between least poor and most poor families, and inequalities in all measured child health outcomes. Overall, 75 (8.5%) infants died between birth and 36 weeks. Infant mortality and HIV transmission was higher among the poorest families within the sample. Immunization coverage was higher among the least poor. The inequalities were mainly due to the area of residence and socioeconomic position. Their study proved that, socio-economic inequalities are highly prevalent within the relatively poor black population in South Africa. Poor socio-economic position exposes infants to ill health. In addition, the use of immunization services was lower in the poor households. These inequalities need to be explicitly addressed in future program planning to improve child health for all South Africans.

(Rahman and Sarkar, 2009) in "Determinants of Infant and Child Mortality in Bangladesh" investigated the mortality of children under five years old using information from women's birth histories pertaining to children born during the 10 years before the $1999-2000$ Bangladesh Demographic and Health Survey. The work specifically provided information on the levels, trends and differentials in neonatal, post-neonatal, infant and child mortality and assessed the effects of socio-economic, demographic and mother's health-care characteristics on infant and child mortality. They concluded that residence, education of father and mother, preceding birth interval, family size, toilet facility, delivery place and antenatal care are the major contributors of infant and child mortality.

(SEÇKIN, 2009) in his thesis "Determinants of infant mortality in Turkey" , examines regional, household and individual level characteristics that are associated with infant mortality. For this purpose survival analysis was used in his analysis. He used data from $2003-2004$ Turkey Demographic and Health Survey that includes detailed information of 8,075 ever married women between the ages $15-49$. 7,360 mothers of these women gave birth to 22,443 children. The results of the logistic regression show that intervals between the births of the infants are associated with infant mortality at lower levels of wealth index. Children from poorer families with preceding birth interval shorter than 14 months or children whose mothers experience a subsequent birth fare badly. Breastfeeding was important for the survival chance of the infants under the age 3 months. Place of delivery and source of water the family uses were also found to be correlated with infant mortality risk. Curvilinear relation between maternal age at birth and infant mortality risk was observed, indicating higher risk for teenage mothers and mother's having children at older ages.

(Maydana et al., 2009) did a study with an objective to evaluate socioeconomic

inequalities and its relation to infant mortality in Bolivia's municipalities in 2001. An ecological study based on data from the 2001 National Census on Population and Housing covering the 327 municipalities in Bolivia's nine departments. The dependent variable was the infant mortality rate (IMR); the independent variables were indirect socioeconomic indicators (the percentage of illiterates older than 15 years of age, building materials and sanitation features of the houses). The geographic distribution of each indicator was determined and the associations between IMR and each socioeconomic indicator were calculated using Spearman's rank correlation coefficient and adjusted with Poisson regression models. The resulting IMR for Bolivia in 2001 was 67 per 1,000 live births. Rates ranged from < 0.1 per $1,000$ live births in the Magdalena municipality, Beni department, to 170.0 per 1,000 live births in the Caripuyo municipality, PotosÃ department. The mean rate of illiteracy per municipality was 17.5%; the mean percentage of houses without running water was 90.4%, and for those lacking sanitation services, 67.6%. The IMR was inversely associated with all of the socioeconomic indicators studied. The highest relative risk was found in housing without sanitation services. Multi factorial models adjusted for illiteracy showed that the following indicators were still strongly associated with the IMR: no sanitation services (Relative risk (RR) = 1.54; 95% Confidence Interval $(95\%CI) = 1.38 - 1.66$); adobe, stone, or mud walls $(RR = 1.54; 95\%CI : 1.43 - 1.67)$; and, corrugated metal, straw, or palm branch roof $(RR = 1.34; 95\%CI : 1.26 - 1.43)$. A significant association was found between poor socioeconomic status and high IMR in Bolivia's municipalities in 2001. The municipalities in the country's central and southeastern areas had lower socioeconomic status and higher IMR. The lack of education, absence of basic sanitation and precarious housing conditions were key factors that tripled the risk of death.

(Kraft et al., 2013) conducted a study which uses data from the Philippines to assess trends in the prevalence and distribution of child mortality and to evaluate

the country's socioeconomic-related child health inequality using the Philippines as a case study. Using data from four Demographic and Health Surveys they estimated levels and trends of neonatal, infant, and under-five mortality from 1990 to 2007. Mortality estimates at national and sub national levels were produced using both direct and indirect methods. Concentration indices were computed to measure child health inequality by wealth status. Multivariate regression analyses were used to assess the contribution of interventions and socioeconomic factors to wealth-related inequality. The paper shows that, despite substantial reductions in national under-five and infant mortality rates in the early 1990s, the rate of declines have slowed in recent years and neonatal mortality rates remain stubbornly high. Substantial variations across urban-rural, regional and wealth equity-markers were evident, and suggest that the gaps between the best and worst performing sub-populations will either be maintained or widen in the future. Of the variables tested, recent wealth-related inequalities were found to be strongly associated with social factors (e.g. maternal education), regional location and access to health services such as facility-based delivery. It was concluded that, the Philippines has achieved substantial progress towards Millennium Development Goal 4, but this success masks substantial inequalities and stagnating neonatal mortality trends. This analysis supports a focus on health interventions of high quality,that is, not just facility-based delivery, but delivery by trained staff at well-functioning facilities and supported by a strong referral system-to re-start the long term decline in neonatal mortality and to reduce persistent within-country inequalities in child health.

(Malderen et al., 2013a) conducted a study with the main objective of decomposing wealth-related inequalities in skilled birth attendance and immunization into their contributing factors. Data from the Kenyan DHS (2008/09) was used. The study investigated the effects of socio-economic determinants on coverage and wealth-related inequalities of skilled birth attendance utilization and measles

immunization. Techniques used were multivariate logistic regression and decomposition of the concentration index (C). The results indicates that skilled birth attendance utilization and measles immunization coverage differed according to household wealth, parent's education, skilled antenatal care visits, birth order and father's occupation. Skilled birth attendance utilization further differed across provinces and ethnic groups. The overall C for skilled birth attendance was 0.14 and was mostly explained by wealth (40%), parent's education (28%), antenatal care (9%) and province (6%). The overall C for measles immunization was 0.08 and was mostly explained by wealth (60%), birth order (33%), and parent's education (28%). Rural residence (−19%) reduced this inequality. It was concluded that both health care indicators require a broad strengthening of health systems with a special focus on disadvantaged sub-groups.

(Malderen et al., 2013b) conducted another study of which the purpose was to investigate and compare the main determinants of overall inequality and wealth-related inequality in under-5 mortality in 13 African countries. Data were from Demographic and Health Surveys conducted in $2007-2010$ in African countries. The study assessed the contribution of determinants to overall inequality in under-5 mortality measured by the Gini index and wealth related inequality in under-5 mortality measured by the concentration index. Techniques used were multivariate logistic regression and decomposition of Gini and concentration indexes (C). Normalized $C(C^*)$ proposed by Erreygers was computed when comparing countries. The results indicated that birth order, birth interval and region contributed the most to overall inequality in under-5 mortality in a majority of countries. A significant wealth-related inequality was observed in five countries: DR Congo, Egypt, Madagascar, Nigeria and Sao Tome and Principe. Under-5 mortality ranged from 2.48% in Egypt to 11.14% in Nigeria. Across all countries, the median under-5 mortality was 8.25% (Interquartile range: $6.12\%-10.29\%$). Among countries with the lowest under-5 mortality, Sao Tome & Principe had the highest level

of inequality compared with Egypt and Madagascar ($C* = -0.032(-0.052to -0.013)$, -0.012 $(-0.018to - 0.005)$ and -0.013 $(-0.022to - 0.004)$, respectively). Among countries with the highest under-5 mortality, DR Congo and Nigeria had the highest level of inequality compared with Sierra Leone and Lesotho ($C^* = -0.050(-0.065 - -0.035), -0.057(-0.065 - -0.050), -0.007(-0.027 - 0.013)$ and $-0.018(-0.038to0.001)$, respectively). Overall, household wealth, father's occupation and mother's education contributed the most to this inequality, though the ranking of the most important determinants differed across countries. It was concluded that assessing the contribution of determinants to overall inequality and to wealth-related inequality in under-5 mortality help in prioritizing interventions aiming at improving child survival and equity.

(S et al., 2012) assessed the health system performance in their paper titled "Decomposing socioeconomic inequality in self-rated health in Tehran" Self-rated health (SRH) and demographic characteristics, including gender, age, marital status, educational years and assets were measured by structured interviews of 2464 residents of Tehran in 2008. A concentration index was calculated to measure health inequality by economic status. The association of potential determinants and SRH was assessed through multivariate logistic regression. The contribution to concentration index of level of education, marital status and other determining factors was assessed by decomposition. They found out that, the mean age of respondents was 41.4 years of which 49% of them were men. The mean score of SRH status was 3.72. 282 respondents (11.5%) rated their health status as poor or very poor. The concentration index was -0.29. Age, marital status, level of education and household economic status were significantly associated with SRH in both the crude and adjusted analyses. The main contributors to inequality in SRH were economic status (47.8%), level of education (29.2%) and age (23.0%). They concluded that, sub-optimal SRH was more in lower economic status than in higher economic status.

(Mosquera et al., 2012) did a study on the impact of primary health care in reducing inequalities in child health outcomes in Bogota, the capital of Colombia. A Primary Health Care (PHC) strategy was implemented in 2004 to improve health care and to address the social determinants of such inequalities. The study aimed to evaluate the contribution of the Primary Health Care (PHC) strategy to reducing inequalities in child health outcomes in Bogota. Their study made use of an ecological analysis with localities as the unit of analysis. The variable used to capture the socioeconomic status and living standards was the Quality of Life Index (QLI). Concentration curves and concentration indices for four child health outcomes (infant mortality rate (IMR), under-5 mortality rate, prevalence of acute malnutrition in children under-5 and vaccination coverage for diphtheria, pertussis and tetanus) were calculated to measure socioeconomic inequality. Two periods were used to describe possible changes in the magnitude of the inequalities related with the PHC implementation (2003 year before-2007 year after implementation). The contribution of the PHC intervention was computed by a decomposition analysis carried out on data from 2007. The outcome of the results showed that in both 2003 and 2007, concentration curves and indexes of IMR, under-5 mortality rate and acute malnutrition showed inequalities to the disadvantage of localities with lower QLI. Diphtheria, pertussis and tetanus (DPT) vaccinations were more prevalent among localities with higher QLI in 2003 but were higher in localities with lower QLI in 2007. The variation of the concentration index between 2003 and 2007 indicated reductions in inequality for all of the indicators in the period after the PHC implementation. In 2007, PHC was associated with a reduction in the effect of the inequality that affected disadvantaged localities in under-5 mortality (24%), IMR (19%) and acute malnutrition (7%). PHC also contributed approximately 20% to inequality in DPT coverage, favoring the poorer localities. It was concluded that, the PHC strategy developed in Bogota appears to be contributing to reductions of the inequality associated

with socioeconomic and living conditions in child health outcomes.

(WAGSTAFF, 2000) titled "socioeconomic inequalities in child mortality: comparison across nine developing countries" tried to analyze data from Brazil, Cote d'Ivoire, Ghana, Nepal, Nicaragua, Pakistan, the Philippines, South Africa and Vietnam. The paper aims to diminish this information gap by outlining methods of measuring health inequalities between the poor and the non-poor and of testing for significant differences across countries or temporal changes. Also to generate evidence on the magnitude of inequalities between the poor and non-poor in a particular dimension of health, namely mortality and in a particular section of the population of the developing world, namely children aged under-five years. The data used was obtained from the living Standards Measurement Study and the Cebu Longitudinal Health and Nutrition Survey. Mortality rates were estimated directly where complete fertility histories were available and indirectly otherwise. Mortality distributions were compared between countries by means of concentration curves and concentration indices. Dominance checks were carried out for all pair wise inter country comparisons. Standard errors were calculated for the concentration indices and tests of inter country differences in inequality were performed. The application of concentration curves and indices to the data showed that inequalities in infant and under-five mortality were to the disadvantage of the worse-off. Inequalities in under-five mortality were especially high in Brazil and rather higher in Nicaragua and the Philippines. They were lower in Cote d'Ivoire, Nepal and South Africa but higher in these countries than in Ghana, Pakistan and Vietnam. The results suggested that, for the most part, inequalities in infant and under five mortality favour the better-off, and that these inequalities vary between countries. However, nothing has been said about why inequalities favour the better-off, why they are higher in some countries than in others and what policies might be most cost-effective in reducing them. He then concluded that these matters deserve special attention in future work.

(JP et al., 2008) in their article socioeconomic inequalities in health in 22 European countries compared the magnitude of inequalities in mortality and self-assessed health among 22 countries in all parts of Europe. This comparison among countries can help to identify opportunities for the reduction of inequalities in health. They obtained data on mortality according to education level and occupational class from census-based mortality studies. Deaths were classified according to cause, including common causes, such as cardiovascular disease and cancer; causes related to smoking; causes related to alcohol use; and causes amenable to medical intervention, such as tuberculosis and hypertension. Data on self-assessed health, smoking, and obesity according to education and income were obtained from health or multipurpose surveys. For each country, the association between socioeconomic status and health outcomes was measured with the use of regression-based inequality indexes. The results indicated that, in almost all countries, the rates of death and poorer self-assessments of health were substantially higher in groups of lower socioeconomic status, but the magnitude of the inequalities between groups of higher and lower socioeconomic status was much larger in some countries than in others. Inequalities in mortality were small in some southern European countries and very large in most countries in the eastern and Baltic regions. These variations among countries appeared to be attributable in part to causes of death related to smoking or alcohol use or amenable to medical intervention. The magnitude of inequalities in self-assessed health also varied substantially among countries, but in a different pattern. It was concluded that there are variations across Europe in the magnitude of inequalities in health associated with socioeconomic status. These inequalities might be reduced by improving educational opportunities, income distribution, health-related behavior, or access to health care.

(Zere et al., 2012) studied "Inequities in maternal and child health outcomes

and interventions in Ghana" examined the equity dimension of child and maternal health outcomes and interventions using Ghana as a case study. Data from Ghana Demographic and Health Survey 2008 report was analyzed for inequities in selected maternal and child health outcomes and interventions using population-weighted, regression-based measures: slope index of inequality and relative index of inequality. The results indicated that no statistically significant inequities were observed in infant and under-five mortality, perinatal mortality, wasting and acute respiratory infection in children. However, stunting, underweight in under-five children, anaemia in children and women, childhood diarrhoea and underweight in women BMI(< 18.5) show inequities that are to the disadvantage of the poorest. The rates significantly decrease among the wealthiest quintile as compared to the poorest. In contrast, overweight BMI($25-29.9$) and obesity BMI (≥ 30) among women reveals a different trend-there are inequities in favour of the poorest. In other words, in Ghana overweight and obesity increase significantly among women in the wealthiest quintile compared to the poorest. With respect to interventions: treatments of diarrhoea in children, receiving all basic vaccines among children and sleeping under insecticide treated mosquito net (children and pregnant women) have no wealth-related gradient. Skilled care at birth, deliveries in a health facility (both public and private), caesarean section, use of modern contraceptives and intermittent preventive treatment for malaria during pregnancy all indicate gradients that are in favour of the wealthiest. The poorest use less of these interventions. Not unexpectedly, there is more use of home delivery among women of the poorest quintile. It was concluded that significant Inequities were observed in many of the selected child and maternal health outcomes and interventions. Failure to address these inequities vigorously is likely to lead to non achievement of the MDG targets related to improving child and maternal health (MDGs 4 and 5). The government should therefore give due attention to tackling inequities in health outcomes and use of interventions by implementing equity-enhancing measure both within and outside the health sector in line with

the principles of Primary Health Care and the recommendations of the WHO Commission on Social Determinants of Health.

2.1 other related literatures

(P et al., 2006) in their paper titled "using classification trees to assess low birth weight outcomes" used classification trees to study the interactive nature of risk factors. In particular they identify subgroups of women who are at a high risk of a low birth weight (LBW) outcome in seven geographical regions of Florida, and study the predictive performance of classification trees by comparing the tree-based results to those obtained using logistic regression. The data, 181,690 singleton births, were derived from Florida birth certificates recorded in 1998. Classification trees and logistic regression models were built based on seven geographical regions. The outcome variable consisted of two classes, namely LBW ($< 2500g$) and normal birth weight ($2500g$) cases, while a large number of known risk factors was examined. Tree and logistic regression models were compared using Receiving Operating Curves, and sensitivity and specificity analyses. The use of classification trees revealed a number of high-risk subgroups. For instance, White, Hispanic or Other non-white mothers who were healthy and smoked with a weight gain less than 20 lbs had a higher risk of a LBW birth compared to those with the same characteristics but with a weight gain of more than 20 lbs. Factors such as parity and marital status were important predictors for pregnancy outcomes among nonsmoker White, Hispanic or Other non-white. Furthermore, they found that Black mothers were directly classified as a high-risk subgroup in the regions of Panhandle, Northeast, North Central, while in the Southern regions a series of other characteristics further defined the high-risk subgroup of Black mothers. Overall, the differences in predictive performance between tree models and logistic regression were minimal. Their study demonstrated that classification trees can be used to identify high-risk subgroups of mothers who are at risk of LBW outcomes. Although these exploratory tree analyses revealed a number

of distinctive variable interactions for each geographical area, the variable selection was similar across all seven regions. Their study also demonstrated that classification trees did not outperform logistic regression models or vice versa; both approaches provided useful analyses of the data.

(Nagy et al., 2010) in their paper "Tree-Based Methods as an Alternative to Logistic Regression in Revealing Risk Factors of Crib-Biting in Horses" tried to establish the difference between tree-based methods and logistic regression. An important difference between these two statistical approaches is that logistic regression makes a number of assumptions about the underlying data, whereas tree-based methods do not. Another difference is that logistic regression can be used to derive odds ratios for the significant risk factors, whereas tree-based methods create a tree where the ramifications represent the risk factors. The probability of occurrence is assigned to each end of branch in the tree. Data of horses used for non-competition purposes were analyzed with three statistical approaches: logistic regression, classification tree, and conditional inference tree methods. By this, they compared the advantages and disadvantages of these statistical methods. No difference was found between the two tree-based methods regarding the structure and prediction accuracy of the trees. Compared to them, logistic regression revealed fewer risk factors, and also the number of the stereotypic horses classified correctly by the model was less. The representation of the tree-based methods is closer to medical reasoning and also high-order interaction of the risk-factors can easily be visualized. Their results suggest that tree-based methods can be a new alternative in revealing risk factors, even if used alone or together with logistic regression.

In addition, (Kajungu et al., 2012) used classification tree modeling to investigate drug prescription practices at health facilities in rural Tanzania. The aim of their study was to understand the factors influencing prescription patterns and

to develop strategies to mitigate the negative consequences associated with poor practices in both the public and private sectors. A cross-sectional study was conducted in rural Tanzania among patients attending health facilities, and health workers. Patients, health workers and health facilities-related factors with the potential to influence drug prescription patterns were used to build a model of key predictors. Standard data mining methodology of classification tree analysis was used to define the importance of the different factors on prescription patterns. Their analysis included 1,470 patients and 71 health workers practicing in 30 health facilities. Patients were mostly treated in dispensaries. Twenty two variables were used to construct two classification tree models: one for polypharmacy (prescription of ≥ 3 drugs) on a single clinic visit and one for co-prescription of artemether-lumefantrine (AL) with antibiotics. The most important predictor of polypharmacy was the diagnosis of several illnesses. Polypharmacy was also associated with little or no supervision of the health workers, administration of AL and private facilities. Co-prescription of AL with antibiotics was more frequent in children under five years of age and the other important predictors were transmission season, mode of diagnosis and the location of the health facility. It was concluded that Standard data mining methodology is an easy-to-implement analytical approach that can be useful for decision-making. Polypharmacy is mainly due to the diagnosis of multiple illnesses.

Lastly, a study was carried out by (Austin, 2008) to compare classification trees grown using R with those grown using S-PLUS. R and S-PLUS are two statistical programming languages that share a similar syntax and functionality. Both R and S-PLUS allow users to fit classification and regression trees. Using data on 9,484 patients hospitalized with an acute myocardial infarction, they compared the classification trees for predicting mortality that were grown using R and S-PLUS. They also used repeated split-sample derivation to determine the predictive accuracy of classification trees grown using R and S-PLUS. Their findings

show that classification tree grown using R was substantially more parsimonious than the one grown using S-PLUS. The pruned classification tree grown using R was equal to a classification tree that was obtained by removing six sub trees from the pruned classification tree grown using S-PLUS. Repeated split-sample validation was then used to demonstrate that classification trees constructed using S-PLUS had greater discrimination and accuracy compared to classification trees grown using R. It was then concluded that R can produce different classification trees than S-PLUS using the same data.

Chapter 3

Methodology

3.1 Introduction

Due to the advancement of technology especially in the field of computer science, one can easily make analysis of data with little knowledge about the statistical and mathematical concepts that underline it. It is important that one acquires the best of knowledge and understanding about the theoretical and conceptual framework of the statistical method that is used to analyze the data efficiently and effectively. Therefore this chapter focuses on the theoretical and conceptual frame work of classification trees which is the basic method used in this research work.

3.1.1 Data description

The data utilized for this analysis was derived from the GDHS conducted in 2008. The data makes use of a two-stage sample design. The first stage involves selecting sample clusters from an updated master sampling constructed from the 2000 Ghana population and housing census. A total of 412 clusters were selected using systematic sampling with probability proportional to size. The second stage involves the systematic sampling of 30 of the households listed in each cluster. A total of 12,323 households were selected of which 11,378 were successfully interviewed yielding a response rate of 99%. Three questionnaires were used, the household questionnaire, the women's questionnaire and the men's questionnaire. The household questionnaire was used to record all deaths of household members that occurred since January 2003, including child mortality. Based on this information, in each household that reported the death of a child under age five years

since 2005, field editors administered a Verbal Autopsy Questionnaire.

3.1.2 Classification trees

(Breiman L. and Stone, 1984) developed classification and regression tree (CART) which is a sophisticated program for fitting trees to data. CART analysis is a tree-building technique which is different from traditional data analysis methods. In a number of studies CART has been found to be quite effective for creating decision rules which perform well or better than rules developed using more traditional methods. Classification tree methods such as CART are convenient way to produce a prediction rule from a set of observations described in terms of a vector of features and a response value. The main aim is to define a general prediction rule which can be used to assign a response value to the cases solely on the bases of their explanatory variables. An attractive feature of the CART methodology is that because the algorithm asks a sequence of hierarchical questions, it is relatively simple to understand and interpret the results.

A classification tree is the result of asking an ordered sequence of questions, and the type of question asked at each step in the sequence depends upon the answers to the previous questions of the sequence. The sequence terminates in a prediction of the class. The unique starting point of a classification tree is called the root node and consists of the entire learning set L at the top of the tree. A node is a subset of the set of variables, and it can be a terminal or non terminal node. A non terminal (or parent) node is a node that splits into two daughter nodes (a binary split). Such a binary split is determined by a Boolean condition on the value of a single variable, where the condition is either satisfied ("yes") or not satisfied ("no") by the observed value of that variable. All observations in L that have reached a particular (parent) node and satisfy the condition for that variable drop down to one of the two daughter nodes; the remaining observations at that (parent) node that do not satisfy the condition drop down to the other daughter

node. A node that does not split is called a terminal node and is assigned a class label. Each observation in L falls into one of the terminal nodes. When an observation of unknown class is "dropped down" the tree and ends up at a terminal node, it is assigned the class corresponding to the class label attached to that node. There may be more than one terminal node with the same class label. A single-split tree with only two terminal nodes is called a stump. The set of all terminal nodes is called a partition of the data.

Example; given a recursive partitioning involving two input variables,X_1 and X_2. Consider the tree diagram in Figure 3.1. The possible stages of this tree are as follows:

1. Is $X_2 \leq \theta_1$? If the answer is yes, follow the left branch; if no, follow the right branch.

2. If the answer to 1 is yes, then we ask the next question, Is $X_1 \leq \theta_2$? An answer of yes yields terminal node τ_1 with corresponding region $R_1 = \{X_1 \leq \theta_2, X_2 \leq \theta_1\}$; an answer of no yields terminal node τ_2 with corresponding region $R_2 = \{X_1 > \theta_2, X_2 \leq \theta_1\}$.

3. If the answer to 1 is no, we ask the next question: Is $X_2 \leq \theta_3$? If the answer to 3 is yes, then we ask the next question: Is $X_1 \leq \theta_4$? An answer of yes yields terminal node τ_3 with corresponding region $R_3 = \{X_1 \leq \theta_4, \theta_1 < X_2 \leq \theta_3\}$; if no, follow the right branch to terminal node τ_4 with corresponding region $R_4 = \{X_1 > \theta_4, \theta_1 \leq X_2 < \theta_3\}$;

4. If the answer to 3 is no, we arrive at terminal node τ_5 with corresponding region $R_5 = \{X_2 > \theta_3\}$.

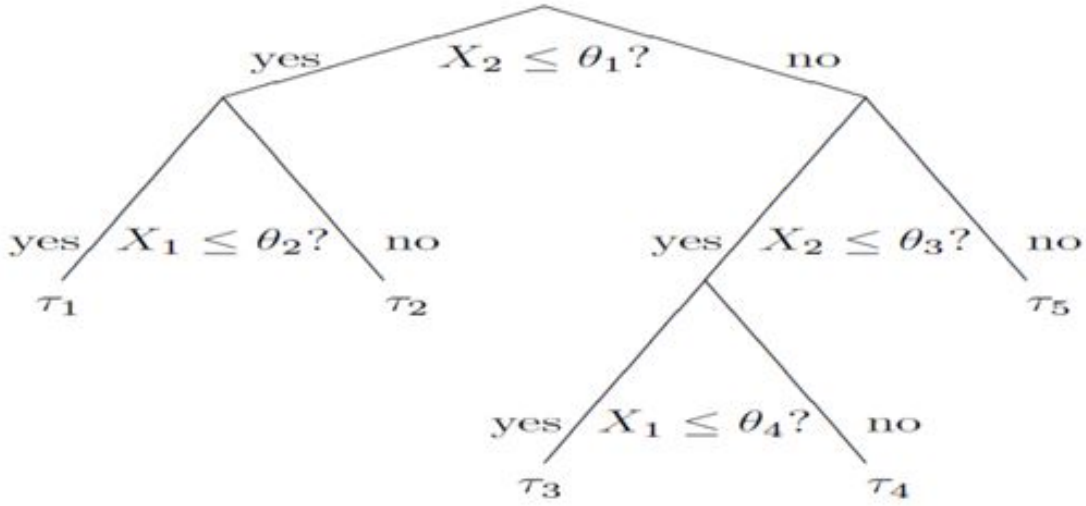


FIGURE 3.1: decision tree with five terminal nodes, $\tau_1 - \tau_5$

The underlining assumption is that $\theta_2 < \theta_4$ and $\theta_1 < \theta_3$. The resulting 5-region partition of R^2 is given in Figure 3.2. For a classification tree, each terminal node and corresponding region is assigned a class label.

FIGURE 3.2: partitioning of $\mathcal{R}^2$ into five regions, $R_1 - R_5$, corresponding to the five terminal nodes.

3.1.3 Tree growing procedure

In order to grow a classification tree, there are four basic questions we need to answer:

- How do we choose the binary conditions for splitting at each node?
- Which procedure should we use to split a parent node into its two daughter

nodes?

- How do we decide when a node become a terminal node (that is, stop splitting)?
- How do we assign a class to a terminal node?

3.1.4 Splitting strategies

In determining how to divide subsets L to create two daughter nodes from a parent node, the tree-growing algorithm at each node has to decide on which variable it is "pure" to split. We need to consider every possible split over all variables present at that node, then enumerate all possible splits, evaluate each one, and decide which is best in some sense. For a continuous or ordinal variable, the number of possible splits at a given node is one fewer than the number of its distinctly observed values. For an equal interval predictor with m distinct values, there are $m-1$ splits that maintain the existing ordering of values. So, $m-1$ splits on that variable need to be evaluated. For example, if there are 40 distinct cumulated weighted average score possible for students, there are 49 possible splits that maintain the existing order.((Breiman L. and Stone, 1984)).

Suppose that a particular categorical variable is defined by M distinct categories, $\{l_1, l_2, l_3, \cdots, l_M\}$. The set S of possible splits at that node for that variable is the set of all subsets of $\{l_1, l_2, l_3, \cdots, l_M\}$. Denote by τ_L and τ_R the left daughter-node and right daughter-node, respectively, emanating from a (parent) node τ. If we let $M = 3$, then there are $2^M - 2 = 6$ possible splits (ignoring splits where one of the daughter-nodes is empty). However, half of those splits are redundant; for example, the split $\tau_L = \{l_1\}$ and $\tau_R = \{l_2, l_3\}$ is the reverse of the split $\tau_L = \{l_2, l_3\}$ and $\tau_R = \{l_1\}$. So, the set S of six distinct splits is given by Table 3.1:

In general, there are $2^{M-1} - 1$ distinct splits in S for an M-categorical variable.

Table 3.1: *Categorical variable with six distinct splits*

τ_L	τ_R
l_1	l_2 l_3
l_2	l_2 l_3
l_3	l_1 l_2

So in general to get the total number of splits for a classification tree, we add the total number of continuous variables and that of the categorical variables.

Node impurity functions

At each stage of recursive partitioning, all of the allowable ways of splitting a subset of L are considered. The one which leads to the greatest increase in node purity is chosen. This can be achieved using "impurity function", which is a function of the proportion of the learning sample belonging to the possible classes of the response variable. To choose the best split over all variables, we first need to choose the best split over a given variable. The aim is to have as little impurity overall as possible. Accordingly, the "best" split is the one that reduces impurity the most. Let $\pi_1, \pi_2, \pi_3, \cdots, \pi_K$ be the $K \geq 2$ classes. For node τ, we define the node impurity function $i(\tau)$ as

$$i(\tau) \quad = \quad \phi(p(1|\tau), p(2|\tau), \cdots, p(K|\tau)) \tag{3.1}$$

Where $p(K|\tau)$ is an estimate of $p(X \in \pi_K|\tau)$,the conditional probability that an observation X is in π_K given that it falls into node τ. In $i(\tau)$ it is important for ϕ to be symmetric functions, defined on the set of all K- tupples of probabilities $(p_1, p_2, \cdots, p_K)$ with unit sum,minimized at the points $(1, 0, 0, \cdots, 0)$, $(0, 1, 0, 0, \cdots, 0)$, $\cdots$, $(0, 0, 0, \cdots, 1)$ and maximize at the point $(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K})$. Assuming $K = 2$, then these conditions reduces to a symmetric $\phi(p)$ maximized at the point $p = \frac{1}{2}$, with $\phi(0) = \phi(1) = 0$

There remains the need to define ϕ . Two of such functions are the Entropy

function and the Gini diversity index. The entropy function is defined as

$$i(\tau) = -\sum_{k=1}^{K} p(k|\tau) \log p(k|\tau) \tag{3.2}$$

When $k = 2$, the entropy function will be,

$$i(\tau) = -\sum_{k=1}^{2} p(k|\tau) \log p(k|\tau) \tag{3.3}$$

$$\begin{aligned} i(\tau) &= -(p(1|\tau) \log p(1|\tau) + p(2|\tau) \log p(2|\tau)) \\ Let p &= p(1|\tau) then 1 - p = p(2|\tau) \\ i(\tau) &= -p \log p - (1-p) \log(1-p) \end{aligned} \tag{3.4}$$

Using the Gini diversity Index for ϕ, we have

$$i(\tau) = \sum_{k \neq k'} p(k|\tau) p(k'|\tau) = 1 - \sum_{k} (p(k|\tau))^2 \tag{3.5}$$

$$i(\tau) = p(1|\tau)p(2|\tau) + p(2|\tau)p(1|\tau) = 2p(1-p) \tag{3.6}$$

These two functions are concave, having minimum at $p = 0$ and $p = 1$ and a maximum at $p = 0.5$. The Gini Index is more likely to partition the data so that there is one relatively homogeneous node having relatively few cases. Entropy tends to partition the data so that all of the nodes for a given split are about equal in size and homogeneity. Practically, there is not much difference between these two types of node impurity functions.

Choosing the best split for a variable

Suppose, at node τ , we apply split s so that a proportion p_L of the observations drops down to the left daughter-node τ_L and the remaining proportion p_R drops down to the right daughter-node τ_R. For example, suppose we have a data set

in which the response variable Y has two possible outcomes, "yes" and "no" . Suppose that one of the possible splits of the input variable X_j is $X_j \leq c$ vs $X_j > c$, where c is some value of X_j. Then, the 2×2 table can be prepared as shown in Table 3.2. Estimate for $p_L = \frac{n_{\bullet 1}}{n_{\bullet\bullet}}$ and that of $p_R = \frac{n_{\bullet 2}}{n_{\bullet\bullet}}$

Table 3.2: *Two-by-two table for a split on the variable* X_j

	yes	No	Row total
$X_j \leq c$	n_{11}	n_{12}	$n_{1\bullet}$
$X_j > c$	n_{21}	n_{22}	$n_{2\bullet}$
column total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

Using the Entropy function as our measure of impurity, the estimated impurity function becomes

$$i(\tau) = -\left(\frac{n_{\bullet 1}}{n_{\bullet\bullet}}\right) \log\left(\frac{n_{\bullet 1}}{n_{\bullet\bullet}}\right) - \left(\frac{n_{\bullet 2}}{n_{\bullet\bullet}}\right) \log\left(\frac{n_{\bullet 2}}{n_{\bullet\bullet}}\right) \tag{3.7}$$

From Table 3.2, for $X_j \leq c$, we estimate for $p_L = \frac{n_{11}}{n_{1\bullet}}$ and that of $p_R = \frac{n_{12}}{n_{1\bullet}}$ and for $X_j > c$, we estimate $p_L = \frac{n_{21}}{n_{2\bullet}}$ and that of $p_R = \frac{n_{22}}{n_{2\bullet}}$. We then compute the estimates for the daughter nodes τ_L and τ_R

$$i(\tau_L) = -\left(\frac{n_{11}}{n_{1\bullet}}\right) \log\left(\frac{n_{11}}{n_{1\bullet}}\right) - \left(\frac{n_{12}}{n_{1\bullet}}\right) \log\left(\frac{n_{12}}{n_{1\bullet}}\right) \tag{3.8}$$

$$i(\tau_R) = -\left(\frac{n_{21}}{n_{2\bullet}}\right) \log\left(\frac{n_{21}}{n_{2\bullet}}\right) - \left(\frac{n_{22}}{n_{2\bullet}}\right) \log\left(\frac{n_{22}}{n_{2\bullet}}\right) \tag{3.9}$$

The goodness of split s at node τ is given by the reduction in impurity gained by splitting the parent node τ into its daughter nodes, τ_L and τ_R

$$\Delta i(s, \tau) = i(\tau) - \left(\frac{n_{1\bullet}}{n_{\bullet\bullet}}\right) \tau_L - \left(\frac{n_{2\bullet}}{n_{\bullet\bullet}}\right) \tau_R \tag{3.10}$$

The best split for this single variable X_j is the one that has the largest value of (3.10) over all $s \epsilon S_j$, the set of possible distinct splits for X_j.

After splitting the root (parent) node, we continue to divide its two daughter nodes. The partitioning principle is the same. For example to further divide τ_L or τ_R we repeat the previous portioning process with a minor adjustment. That is, the partition uses fewer variables this time. This implies, the number of splits decreases. The recursive partitioning process may proceed until the tree is saturated in the sense that the offspring nodes subject to further division cannot be split. This happens, for instance, when there is only one subject in a node. Note that the total number of allowable splits for a node drops as we move from one layer to the next. As a result, the number of allowable splits eventually reduces to zero, and the tree cannot be split any further. The saturated tree is usually too large to be useful, because the terminal nodes are so small that we cannot make sensible statistical inference; and this level of detail is rarely scientifically interpretable. One way to counter this type of situation is to restrict the growth of the tree. A minimum size of a node is set a priori. We stop splitting when a node is smaller than the minimum. The choice of the minimum size depends on the sample size (e.g., one percent). In some applications, we may also wish to impose the condition that the resulting daughter nodes have a minimal size (say four subjects) to allow meaningful comparisons.

(Breiman L. and Stone, 1984) argued that depending on the stopping threshold, the partitioning tends to end too soon or too late. Accordingly, they made a fundamental shift by introducing a second step, called pruning. Instead of attempting to stop the partitioning, they propose to let the partitioning continue until it is saturated or nearly so. Beginning with this generally large tree, we prune it from the bottom up. The point is to find a subtree of the saturated tree that is most "predictive" of the outcome and least vulnerable to the noise in the data.

3.1.5 Misclassification cost

In many applications, the tree-based method is used for the purpose of prediction. That is, given the characteristics of a subject, we must predict the outcome of this subject before we know the outcome. For example, in the study of Goldman et al. (1982), physicians in emergency rooms must predict whether a patient with chest pain suffers from a serious heart disease based on the information available within a few hours of admission. To this end, we first classify a node τ to either class 0 (normal) or 1 (abnormal), and we predict the outcome of an individual based on the membership of the node to which the individual belongs. Unfortunately, we always make mistakes in such a classification, because some of the normal subjects will be predicted as diseased and vice versa. To weigh these mistakes, we need to compute an estimate of the within-node misclassification rate. The resubstitution estimate of the misclassification rate $R(\tau)$ of an observation in node τ is

$$r(\tau) = 1 - \max_k p(k|\tau) \tag{3.11}$$

For two class case, $K = 2$

$$r(\tau) = 1 - \max(p, 1-p) = \min(p, 1-p) \tag{3.12}$$

The resubstitution estimate $r(\tau)$ in the two-class case is graphed in Figure 3.3 (blue curve). If $p<\frac{1}{2}$, the resubstitution estimate increases linearly in p, and if $p>\frac{1}{2}$, it decreases linearly in p.

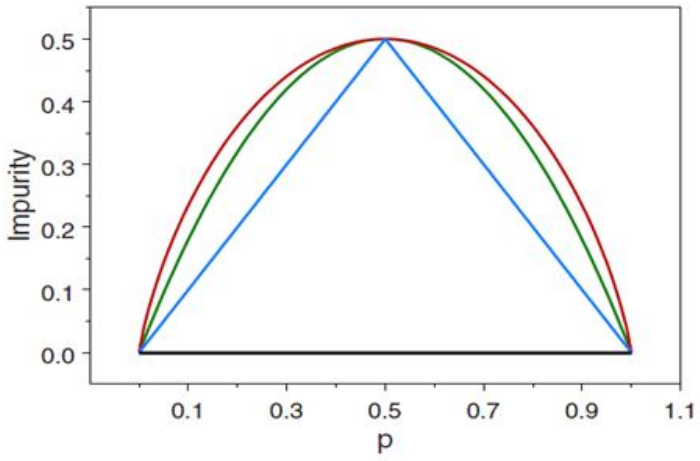


FIGURE 3.30: Node impurity function for two – class case. The entropy function (red), the Gini index (green curve) and the resubstitution estimate of the misclassification rate (blue)

Let T be the tree classifier and $\widetilde{T} = \{\tau_1, \tau_2, \cdots, \tau_L\}$ denote the set of all terminal nodes of T. Then the estimate of the true misclassification rate $R(\tau)$ is

$$R(\tau) = \sum_{\tau \epsilon \widetilde{T}} R(\tau)P(\tau) = \sum_{l=1}^{k} R(\tau_l)P(\tau_l) \tag{3.13}$$

Where $P(\tau)$ is the probability that an observation falls into node τ .If we estimate $P(\tau_l)$ by the proportion $p(\tau_l)$ of all observations that falls into node τ_l , then, the resubstitution estimate of $R(\tau)$ is

$$R^{re}(T) = \sum_{l=1}^{L} r(\tau_l)P(\tau_l) = \sum_{l=1}^{L} R^{re}(\tau_l) \tag{3.14}$$

where $R^{re}(\tau_l) = r(\tau_l)P(\tau_l)$.

The resubstitution estimate $R^{re}(T)$, however, does not work well as an estimate of $R(T)$. First, bigger trees have smaller values of $R^{re}(T)$; that is, $R^{re}(T') \leq R^{re}(T)$, where T' is formed by splitting a terminal node of T. For example, if a tree is allowed to grow until every terminal node contains only a single observation, then that node is classified by the class of that observation and $R^{re}(T) = 0$. Second, using only the resubstitution estimate tends to generate trees that are too big for the given data. Third, the resubstitution estimate $R^{re}(T)$ is a much-too-optimistic estimate of $R(T)$. Hence we introduce the pruning as a real estimate

of $R(T)$

3.1.6 Pruning Tree

Given a classification tree T and an inner node τ. Then pruning of T with respect to τ is the deletion of all successor nodes of τ in T. A pruned tree is a sub tree of the original large tree. How to prune a tree is an important aspect in classification tree. Because there are many different ways to prune a large tree, we decide which is the "best" of those sub trees by using an estimate of $R(T)$.

The pruning algorithm is as follows;

1. Grow the tree, say, T_{max}, where we keep splitting until the nodes each contain fewer than n_{min} observations.

2. Compute an estimate of $R(\tau)$ at each node $\tau \epsilon T_{max}$.

3. Prune T_{max} upwards towards its root node so that at each stage of pruning, the estimate of $R(T)$ is minimized (Izenman, 2008).

Let $\alpha \geq 0$ be a complexity parameter. For any node $\tau \epsilon T$, set

$$R_\alpha(T) = R^{re}(\tau) + \alpha \tag{3.15}$$

We define a cost - complexity pruning measure for a tree as;

$$R_\alpha(T) = \sum_{l=1}^{L} R_\alpha(\tau_l) = R^{re}(T) + \alpha|\widetilde{T}|, \tag{3.16}$$

Where $|\widetilde{T}| = L$ is the number of terminal nodes in the subtree T of T_{max}. $\alpha|\widetilde{T}|$ is considered as a penalty term for tree size, so that $R_\alpha(T)$ penalizes $R^{re}(T)$ for generating too large a tree. For each α, we then choose that subtree $T(\alpha)$ of T_{max} that minimizes $R_\alpha(T)$:

$$R_\alpha(T(\alpha)) = \min_T R_\alpha(T) \tag{3.17}$$

The value α defined on the interval $[0, \infty)$, quantifies the penalty for each additional terminal node. The larger the value of α, the heavier the penalty for complexity. When $\alpha = 0$, there is no penalty and a saturated tree results. So, α is the means by which the size of the tree can be determined. For example, Suppose, for $\alpha = \alpha_1$, the minimizing subtree is $T_1 = T(\alpha_1)$. As we increase the value of α, T_1 continues to be the minimizing subtree until a certain point, say, $\alpha = \alpha_2$, is reached, and a new subtree,$T_2 = T(\alpha_2)$, becomes the minimizing subtree. As we increase α further, the subtree T_2 continues to be the minimizing subtree until a value of α is reached, $\alpha = \alpha_3$, say, when a new subtree $T_3 = T(\alpha_3)$ becomes the minimizing subtree. This argument is repeated a finite number of times to produce a sequence of minimizing subtrees $T_1, T_2, T_3, \cdots$.

How do we get T_{max} to T_1?. Suppose the node τ in the tree T_{max} has daughter node τ_L and τ_R, both of which are terminal nodes. Then,

$$R^{\tau} \geq R^{re}(\tau_L) + R^{re}(\tau_R) \tag{3.18}$$

If equality occurs at node τ, then prune the terminal nodes τ_L and τ_R from the tree. Continue this method until no further pruning of this type is possible. The resulting tree is T_1. Next we find T_2. Let τ be any nonterminal node of T_1, let T_τ be the subtree whose root node is τ , and let $\widetilde{T} = \{T_1^{'}, T_2^{'}, \cdots, T_{L_\tau}^{'}\}$ be the set of terminal nodes of T_τ. Let

$$R^{re}(T_\tau) = \sum_{\tau' \epsilon \widetilde{T_\tau}} R^{re}(\tau') = \sum_{l'=1}^{L_\tau} R^{re}(\tau_{l'}^{'}) \tag{3.19}$$

Then $R^{re}(\tau) > R^{re}(T_\tau)$. Now set

$$R_\alpha(T_\tau) = R^{re}(T_\tau) + \alpha|\widetilde{T_\tau}| \tag{3.20}$$

If $R_\alpha(\tau) > R_\alpha(T_\tau)$, then the subtree T_τ has a smaller cost-complexity than its root node τ, and therefore it is important to maintain T_τ. Substituting (3.15) and

(3.20) into this condition and solving for α yields

$$\alpha < \frac{R^{re}(\tau) - R^{re}(T_\tau)}{\left|\widetilde{T}_\tau\right| - 1} \tag{3.21}$$

For $\tau \epsilon T_1$, define

$$g_1(\tau) = \frac{R^{re}(\tau) - R^{re}(T_{1,\tau})}{\left|\widetilde{T}_{1,\tau}\right| - 1} \tau \notin \widetilde{T}(\alpha_1) \tag{3.22}$$

Where $T_{1,\tau} = T_\tau$

Then, $g_1(\tau)$ can be considered as a critical value for α as long as $g_1(\tau) > \alpha_1$, we do not prune the nonterminal nodes $\tau \epsilon T_1$.

We define the weakest-link node $\widetilde{\tau}$ as the node in T_1 that satisfies

$$g_1(\widetilde{\tau}_1) = \min_{\tau \epsilon T_1} g_1(\tau) \tag{3.23}$$

As α increases, $\widetilde{\tau}_1$ is the first node for which $R_\alpha = R_\alpha(T_\tau)$, so that $\widetilde{\tau}_1$ is preferred to $T_{\widetilde{\tau}_1}$. Set $\alpha_2 = g_1(\widetilde{\tau}_1)$ and define the subtree $T_2 = T(\alpha_2)$ of T_1 by pruning away the subtree $T_{\widetilde{\tau}_1}$ from T_1. Next, to find T_3, we find the weakest-weakest link node $\widetilde{\tau}_2 \epsilon T_2$ through the critical value

$$g_2(\tau) = \frac{R^{re}(\tau) - R^{re}\left(T_{2,\tau}\right)}{\left|\widetilde{T}_{1,\tau}\right| - 1} \tau \notin \widetilde{T}(\alpha_2), \tau \epsilon T(\alpha_2) \tag{3.24}$$

Where $T_{2,\tau}$ is that part of T_τ which is in T_2. We then set

$$\alpha_3 = g_2(\widetilde{\tau}_2) = \min_{\tau \epsilon T_2} g_2(\tau) \tag{3.25}$$

And define the subtree T_3 of T_2 by pruning away the subtree $T_{\widetilde{\tau}_2}$ so that $\widetilde{\tau}_2$ becomes a terminal node from T_2. This process is continued for a number of steps.

Theorem 3.1.1 ((Breiman L. and Stone, 1984)) *For any value of the com-*

plexity parameter α, there is a unique smallest subtree of T_0 that minimizes the cost-complexity

Since there may be several minimizing subtrees for each α ,then , for a given α we call $T(\alpha)$ the smallest minimizing subtree if it is a minimizing subtree and satisfies the following condition:

$$\text{if } R_\alpha(T) = R_\alpha(T(\alpha)) \text{ then } T > T(\alpha)$$

Hence, $T(\alpha)$ is a subtree of T and in the event of any ties, $T(\alpha)$ is taken to be the smallest tree out of all those trees that minimize R_α. This produces a finite increasing sequence of complexity parameters,

$$0 = \alpha_1 < \alpha_2 < \alpha_3 \cdots < \alpha_M$$

Which corresponds to a finite sequence of nested subtrees of T_{max},

$$T_{max} = T_0 > T_1 > T_2 > \cdots > T_M$$

Where $T_k = T(\alpha_k)$ is the unique smallest minimizing subtree for $\alpha \epsilon (\alpha_k, \alpha_{k+1})$, and T_M is the root-node subtree.

Choosing the Best subtree

The sequence of subtrees produced by the pruning procedure serves as the set of candidate subtrees for the model and to obtain the classification tree, all that remains to be done is to select the one which will hopefully have the smallest misclassification rate for future observations. Breiman et al (1984) offered two estimation methods, which is the independent test sample or cross-validation.

Independent Test Set

Independent test is used to estimate the error rates of the various trees in the nested sequence of subtrees, and the tree with minimum estimated misclassification rate can be selected to be used as the tree-structured classification model.

For this purpose, the observations in the learning dataset (D) are randomly assigned to two disjoint datasets, a training dataset (L) and a test set (T), where $L \cap T = \phi$ and $L \cup T = D$. Suppose there are n_τ observations in the test set and that they are drawn independently from the same underlying distributions as the observations in L. Then the tree T_{max} is grown from the learning set only and it is pruned from bottom up to give the sequence of subtrees $T_1 > T_2 > T_3 > \cdots > T_M$, and a class is assigned to each terminal node.

Once a sequence of subtrees has been produced, each of the n_τ test-set observations are dropped down the tree T_k. Each observation in T is then classified into one of the different classes. Because the true class of each observation in T is known, $R(T_k)$ is estimated by $R^{ts}(T_k)$ which is (3.16) with $\alpha = 0$; that is $R^{ts}(T_k) = R^{re}(T_k)$, the resubstitution estimate computed using the independent test set. When the costs of misclassification are identical for each class, $R^{ts}(T_k)$ is the proportion of all test set observations that are misclassified by T_k. These estimates are then used to select the best pruned subtree T^* by the rule,

$$R^{ts}(T*) = \min_k R^{ts}(T_k) \tag{3.26}$$

and $R^(T*)$ is its estimated misclassification rate.

The standard error of $R^{ts}(T)$ is estimated as follows. When test set observations are dropped down the tree T, the chance that any one of these observations are misclassifiied is $p^* = R(T)$. Thus, it is a binomial sampling situation with n_τ Bernoulli trials and probability of success p^*. If $p = R^{ts}(T)$ is the proportion of misclassified observations in T,then p is unbiased for p^* and the variance of p is

$$\frac{p^*(1-p^*)}{n_\tau} \tag{3.27}$$

The standard error of $R^{ts}(T)$ is therefore estimated by

$$\widehat{SE}(R^{ts}(T)) = \left\{ \frac{R^{ts}(1 - R^{ts}(T))}{n_\tau} \right\}^{\frac{1}{2}} \tag{3.28}$$

Cross-validation estimate

For a V-fold cross-validation (CV/V), the learning dataset D is divided into V roughly equal-sized, disjoint subsets,

$$D = \bigcup_{v=1}^{v} D_v \tag{3.29}$$

where $D_v \cap D_{v'} = \emptyset$, $v \neq v'$, and V is taken to be 5 or 10. Next, V different datasets are obtained from the D_v by taking $L_v = D - D_v$ as the v-th training set and $T_v = D_v$ as the v-th test set,$v = 1, 2, 3, \cdots, V$. The v-th tree $T_{max}(v)$ is grown using v-th training set L_v, $v = 1, 2, 3, \cdots, V$. The value of the complexity parameter α is fixed to a certain value. Let, $T^{(v)}(\alpha)$ be the best pruned subtree of $T_{max}(v)$. Now, each observation in the v-th test T_v is dropped down the $T^(v)(\alpha), v = 1, 2, 3, \cdots, V$. Let $n_{ij}^{(v)}$ be the number of j-th class observations in T_v that are classified as being from the i-th class, $i, j = 1, 2, 3, \cdots, K$, $v = 1, 2, 3, \cdots, V$. Because

$$D = \bigcup_{v=1}^{v} T_v \tag{3.30}$$

is a disjoint sum, the total number of j-th class observations that are classified as being from the i-th class is

$$n_{ij} = \sum_{ij}^{(v)} (\alpha), i, j = 1, 2, 3, \cdots, K \tag{3.31}$$

If n_j is the number of observations in D that belong to the j-th class,

$j = 1, 2, 3, \cdots, J$, and assuming equal misclassification cost for all classes, then for a given α,

$$R^{CV/V}(T(\alpha)) = n^{-1} \sum_{i=1}^{K} \sum_{j=1}^{K} n_{ij}(\alpha) \tag{3.32}$$

Is the misclassification rate over D, where $T(\alpha)$ is a minimizing subtree of T_{max}. The final step in this process is to find the right sized subtree. For different values of α, $R^{CV/V}$ is evaluated. If for a sequence of values α_k, corresponding cross-validation error of the minimizing subtree $T(\alpha) = T_k$ is given by,

$$R^{CV/V}(T_k) = R^{CV/V} = (T(\alpha_k)) \tag{3.33}$$

Then, the best-pruned subtree $T*$ is selected by the rule,

$$R^{CV/V}(T*) = \min_k R^{CV/V}(T_k) \tag{3.34}$$

and $R^{CV/V}(T*)$ is used as its estimated misclassification rate.

Chapter 4

Data Analysis and Results

4.1 Data Collection

This chapter deals with exploratory analysis of the data, inferential analysis and modeling based on the study objectives and the information gathered. Data from the Demographic and Health Survey (DHS) conducted in Ghana in 2008 was used for the study. The survey was conducted from 8th September, 2008 to 25th November 2008 on a nationally representative sample of 12,323 households. Each of these households was visited to obtain information about the household using the Household Questionnaire. This was used to identify deaths of children under-five years occurring in the household since January 2005. The sample was selected in such a manner as to allow for separate estimates of key indicators for each of the ten regions in Ghana as well as for urban and rural areas separately.

The study was conducted by Ghana Statistical Service(GSS) and the Ghana Health Service(GHS). Inner City Fund (ICF) Macro, an ICF International Company, provided technical support for the survey through the MEASURE DHS program. Funding for the survey came from the United States Agency for International Development (USAID), through its office in Ghana, and the Government of Ghana, with support from the United Nations Population Fund (UNFPA), the United Nations Children's Fund(UNICEF), the Ghana AIDS Commission (GAC), and the Danish Development Agency (DANIDA).

The 2008 GDHS utilized a two-stage sample design. The first stage involved selecting sample points or clusters from an updated master sampling frame con-

structed from the 2000 Ghana Population and Housing Census. A total of 412 clusters were selected from the master sampling frame. The clusters were selected using systematic sampling with probability proportional to size. A complete household listing operation was conducted from June 2008 to July 2008 in all the selected clusters to provide a sampling frame for the second stage selection of households. The second stage of selection involved the systematic sampling of 30 of the households listed in each cluster. The primary objectives of the second stage of selection were to ensure adequate numbers of completed individual interviews to provide estimates for key indicators with acceptable precision and to provide a sample large enough to identify adequate numbers of under-five deaths to provide data on causes of death. Data were not collected in one of the selected clusters due to security reasons, resulting in a final sample of 12,323 selected households.

The selection criteria used to select ever conceived or ever born mothers in the survey. In this study, sample is limited to 1,295 respondents who gave birth or have children up to 60 months prior to the survey. Even though this is only 27% of the entire sample, the number is large enough for our analysis.

4.2 Exploratory Analysis of Study Variables

The dependent variable in the analysis was under-five mortality. It was evaluated using "yes" for ever experienced mortality child and "no" for not experienced child mortality.

Table 4.1: *Extracted Under-five mortality Data for GDHS* 2008

Mortality	number(%)
No	816 (63.0)
Yes	479 (37.0)
Total	**1295 (100)**

From Table 4.1, the total number of children who survived at the age of five

were 816 which represents 63.0% whilst the number of children who died before attaining the age of five were 479 which represents 37.0% of the children under consideration. This sample represents 1295 mothers who gave birth from the year 2004 to 2008.

The Independent variables considered were categorized into continuous variables and categorical variables. Categorical variables have two main types of measurement scales, that is the nominal and the ordinal scale. The ordinal measurement scales were; highest educational level (HEL) and wealth index (WI) of respondents while the nominal measurement scales includes ethnicity (ETH) of respondents, child born by caesarean (CBC), sex of child (SEX), defacto type of residence (DTOR) and defacto region of residence (DROR). The continuous variables that were considered include age of respondents at birth (AORAB), total number of children ever born (TNOCB), preceding birth interval (PBI). These variables were selected based on related studies example (Hosseinpoor et al., 2006), which suggest that they have predictive power for the question at hand.

The age of respondents at the time of interview and the age of the respondents at birth of a child were categorized into four groups ($15-19$ years; $20-29$ years; $30-39$ years and $40-49$ years). Defacto region of residence by design was classified into ten groups (Ashanti ; Brong Ahafo; Greater Accra ; Volta; Eastern; Western; Central; northern; Upper East; Upper West). Defacto type of residence was divided into two (urban and rural) whilst respondents were asked to report their highest level of education completed, for which responses were grouped as "no education", "primary", "middle" and "higher". Ethnicity was also classified into five and these were "Akan", "Ewe", "Ga/ dangme", "Mole-dagbani" and "others". Wealth index (quintile) was also grouped into poor, middle and rich. The wealth index, computed by GDHS, comprises household assets (type of flooring, water supply, sanitation facilities, electricity, persons per sleeping room,

ownership of agricultural land, domestic servant and other assets).

Table 4.2: *Under-five mortality distribution by socio-economic and demographic characteristics*

Variable	Total(n)	mortality(%)	95% C.I.
Residents			
Urban	368	160(33.4)	(0.113,0.422)
Rural	927	319(66.6)	reference
Region			
Western	83	33(6.9)	reference
Central	110	39(8.1)	(-0.481, 0.355)
Greater Accra	62	35(7.3)	(0.067, 1.169)
Eastern	126	39(8.1)	(-0.644, 0.163)
Volta	80	(28(5.8)	(-0.273, 0.875)
Brong Ahafo	148	41(8.6)	(-0.881, -0.138)
Northern	196	96(20.0)	(-0.003, 0.752)
Upper East	112	39(8.1)	(-0.385, 0.439)
Upper West	135	56(11.7)	(-0.714, 0.201)
Ashanti	233	73(15.2)	(-0.720, 0.075)
Mother's Education			
no education	508	209(43.6)	(-0.091, 0.534)
primary	311	115(24.0)	(-0.104, 0.515)
secondary	454	147(30.7)	reference
higher	22	8(1.7)	(-1.070, 0.369)
Ethnicity			
Akan	524	171(35.7)	(-0.295, 0.275)
Ewe	132	38(7.9)	(-0.998, -0.108)
Ga/Adangme	49	19(4.0)	(-0.646, 0.467)
Mole-dagbani	348	138(28.8)	(-0.117, 0.558)
Others	242	113(23.6)	reference
Wealth quintile			
Poor	508	156(32.6)	(-0.497, 0.155)
Middle	364	130(27.1)	(-0.252, 0.108)
Rich	423	193(40.3)	reference

Table 4.2 shows differentials in childhood mortality by five socio-economic and demographic variables with 95% Wald confidence interval for each variable: residence, region, mother's education, ethnicity and household wealth status (quin-

tile). The observed mortality levels in rural areas are higher than those in urban areas without correction in population densities. The total number of under-five children that were born in the rural area was 927 (71.6%) with observed mortality counts of 319 as compared to that of the urban area with total under-five been 368 (28.4%) and mortality counts of 160 under-fives.

The differences in under-five mortality counts by regions are marked. The under-five mortality counts ranges from a low of 28 observed deaths out of a total of 80 (6.2%) children below the age of five in the Volta region, followed by Western region with under-five mortality count of 33, Greater Accra with under-five mortality count of 35 then Central, Eastern and Upper East regions have under-five mortality counts of 39. Brong Ahafo also recorded an under-five mortality count of 41, Upper West recorded 56 under-five mortality and to a high of 73 and 96 deaths in the Ashanti and Northern regions respectively.

4.3 Statistical Analysis

This aspect of the analysis presents in detailed how the selected statistical techniques have been used to investigate the data and report the findings accordingly. The statistical technique the study deems it suitable for this investigation is classification trees using the R software (Rpart). Detailed computer outputs of this technique would be observed and interpreted.

According to (Adu Brenya,1999), one of the major characteristics of the Ghanaian space economy is the existence of spatial development disparities. He emphasized that disparities exist between the 'North' and the 'South' with the distribution of social and economic infrastructure skewed in favour of the regions in the south; there is also a wide disparity between urban and rural areas; Wide disparities also exist among the districts in all the regions, especially between the rural and urban areas of Ghana.

Test of association between DTOR (urban,rural) and mortality

From table A in appendix, the Pearson chi square test of association reveals that $\chi(1) = 9.290$, $p = 0.002$. This tells us that there is statistically significant association between defacto type of residence (DTOR) and under-five mortality. In view of this, the data was first divided into two (rural and urban) and the factors considered for each. The entire data set was also considered and the factors observed and the inequalities that exist between these two areas are considered. This will help us to explore the socio-economic and demographic differentials that exist.

4.4 Classification Tree

4.4.1 Splitting strategy

At each node, the tree-growing algorithm has to decide on which variable it is "best" to split as discussed in the methodology. In the data, we have four continuous variables; TNOCB (26 possible splits), AORAB (34 possible splits), NOAV (26 possible splits) and PBI (144 possible splits). The total number of possible splits from these continuous variables is therefore, 230. There are also seven categorical variables; HEL (7 possible splits), WI (3 possible splits), ETH (15 possible splits), CBC (1 possible splits) DTOR (1 possible splits), SEX (1 possible split) and DROR (511 possible splits). The total number of possible splits for these categorical splits are therefore 539.

Adding the number of possible splits from the categorical variables (539) to the total number of possible splits from continuous variables (230), we have 769 possible splits over all the 11 variables at the root node. In other words, there are 769 possible splits of the root node into two daughter nodes.

4.4.2 Analysis on Urban

The data consist of 368 observations (160 experienced under-five mortality, 208 has not experience under-five mortality). The classification tree is displayed in Figure 4.1 where we used the entropy measure as the impurity function for splitting (equation 3.3). There are 16 splits and 17 terminal nodes in this tree.

The root node (DROR) with 368 observations is split according to whether DROR = WR, VR, NR, UE, UW (58 observations) or DROR = AR, GA, BA, CR, ER (310 observations). The node with 58 observations was declared a terminal node for "yes" because of $50 - 0$ majority in favour of "yes". The node with the 310 observations which consist of 102 under-five mortality and 208 not experienced mortality, is then split by whether TNOCB < 1.5 (16 observations) or TNOCB ≥ 1 majority in favour of "yes". The node with 294 observations, which consists of 86 experienced mortality and 208 not experienced mortality, is split by whether DROR = GA (39 observations) or DROR = AR, BA, CR, ER (255 observations). The node with 39 observations is then split by whether PBI ≥ 29 (16 observations) or PBI < 29 (23 observations). The node with 16 observations is declared a terminal node because of $16 - 0$ majority in favour of "yes", while that of 23 observations was also declared a terminal node because of $4 - 19$ majority in favour of "no". The node with 255 observations which consist of 66 observations experienced mortality and 189 did not experienced mortality, is split by whether AORAB ≥ 14.5 (247 observations) or AORAB < 14.5 (8 observations). The node with 8 observations is declared terminal node because of $0 - 8$ majority in favour of "no", and the node with 247 observations, which consist of 66 experienced mortality and 181 not experienced mortality is then split by whether PBI < 104 (220 observations) or PBI ≥ 104 (27 observations). The node with 27 observations is declared a terminal node because of $3 - 24$ majority in favour of "no", and the node with 220 observations which consist of 63 experienced mortality and 157 not experienced mortality is then split by whether TNOCB ≥ 3.5 (110 observations)

or TNOCB < 3.5 (110 observations). The rest of the results are shown in Figure 4.1.

Estimating the misclassification rate

From Figure 4.1 and Table 4.3, out of the 368 observations in the data, the classification tree misclassifies 43 of the observations who has experienced under-five mortality (yes) as not experienced it before (no), whereas of the 208 who has not experienced mortality before (no), 12 are misclassified as having experienced mortality before (yes). This gives the resubstitution estimate as

$$R^{re}(T) = \frac{12 + 43}{368} = 0.149 \tag{4.1}$$

Table 4.3: *Misclassification rate for urban*

	Predicted	
Mortality	**Yes**	**No**
Yes	117	43
No	12	196

Pruning the tree

The Breiman et al (1984) philosophy of growing trees is to grow the tree "large" and then prune off branches from the bottom up until the tree is of "right size". A pruned tree is a subtree of the original tree.

Table 4.4: *Cross Parameter table for urban*

CP	**nsplit**	**rel error**	**xerror**$(R^{cv/10})$	**xstd**(T_k)
0.3625000	0	1.00000	1.00000	0.059436
0.1000000	1	0.63750	0.63750	0.053666
0.0500000	2	0.53750	0.53750	0.050738
0.0089286	4	0.43750	0.50000	0.049454
0.0062500	13	0.35625	0.53750	0.050738
0.0031250	14	0.35000	0.58125	0.052103
0.0010000	16	0.34375	0.60625	0.052823

Table 4.4 provides a brief summary of the overall fit of the model. The table is printed from the smallest tree (no splits that is 0) to the largest tree (16 splits)

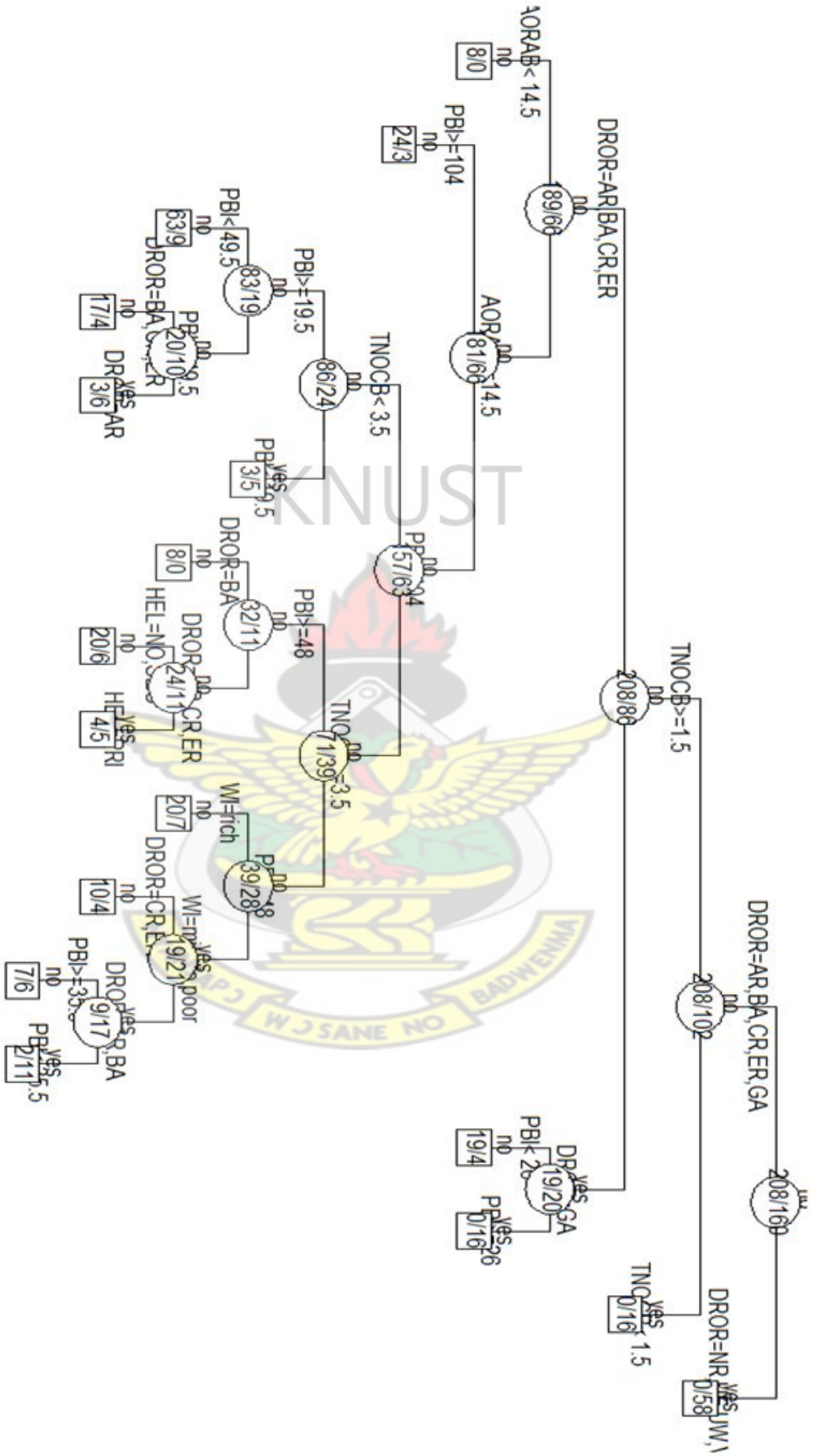


Figure 4.1: *Classification tree for under-five mortality in the urban DHS data*

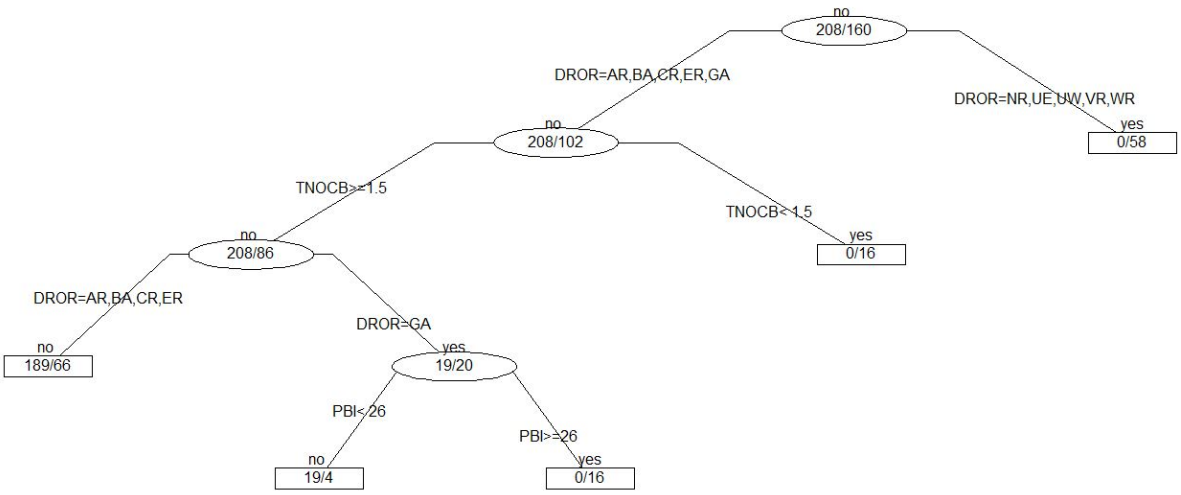


Figure 4.2: *A pruned classification tree with 4 splits and 5 terminal nodes*

and this is shown by the column "nsplit". The number of nodes is always one plus the number of splits. The "CP" column lists the values of the complexity parameter (α) (equation 3.21) at each stage of the tree growing process, and the column "xerror" contains cross-validated classification error rates; the standard deviation of the cross-validation error rates are in the "xstd" column. The relative error (rel error) continues to decrease as the tree becomes more complex, it then begins to rise to some point. This is shown in Figure 4.3. From Table 4.4, the $1-SE$ rule yields a minimum of $CVerror+SE = 0.50000+0.049454 = 0.549454$, which leads to the choice of a classification tree with 4 splits (5 terminal nodes) based upon the cross-validation. The corresponding pruned classification tree is displayed in Figure 4.2.

Variable of importance

Variable of importance describes the role of a variable in a specific tree. It is arranged or ranked from the most important to the less important. According to the discriminatory power in the CART analysis, age of respondents which is the root node emerged as the strongest discriminating factor for under-five mortality, followed by Total number of children ever born (63.70%), Preceding birth interval (59.70%), Ethnicity (40.40%), wealth index (19.33%), Highest educational level

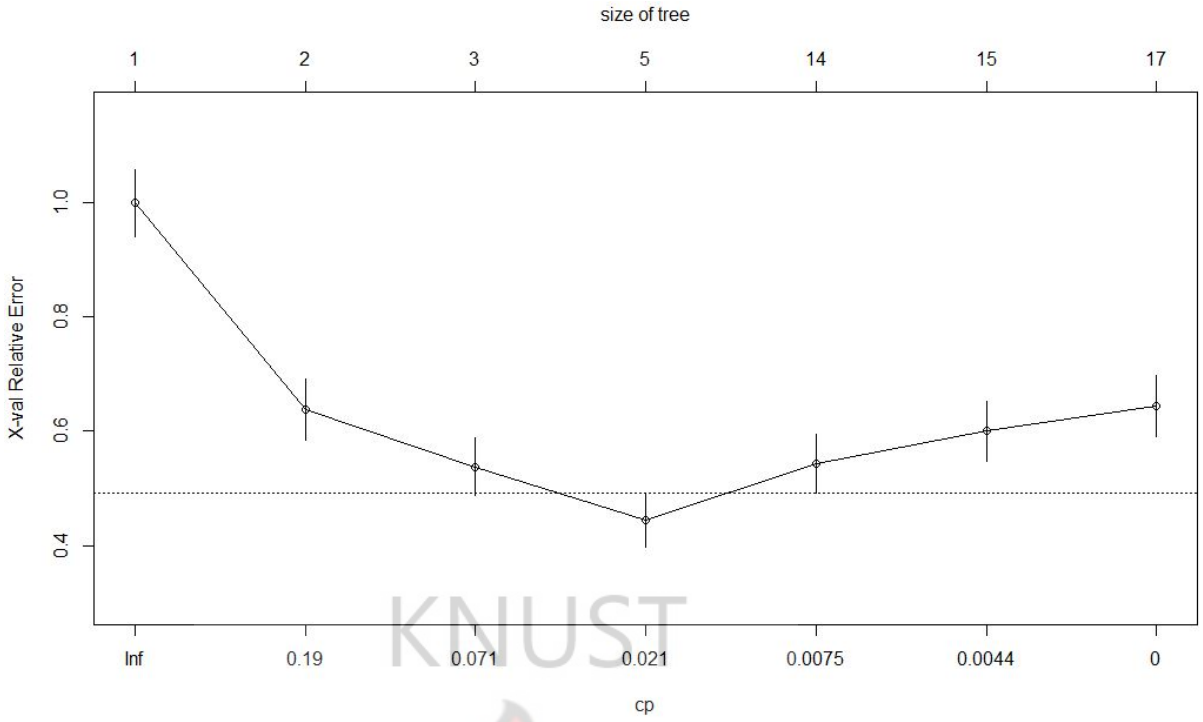


Figure 4.3: *A plot of CP against relative error showing different size of trees for urban*

(18.80%) and sex of child (4.80%) as the least factor that contributes to under-five mortality in Ghana. (Table 4.5).

Table 4.5: *Ranking the predictors of inequalities in under-five mortality in the urban area*

Independent variable	**Power(%)**
Age of respondents at first birth	100.00
Total number of children ever born	63.7
Preceding birth interval	59.70
Region of Residence	48.00
Ethnicity	30.40
Wealth index	19.33
Highest educational level	18.80
Sex of child	4.80

4.4.3 Analysis on Rural

The data consist of 927 (319 experienced under-five mortality, 608 has not experience under-five mortality). The classification tree is displayed in Figure 4.4 where we used the entropy measure as the impurity function for splitting. There

are 27 splits and 28 terminal nodes in this tree

The root node (TNOCB) with 927 observations is split according to whether TNOCB < 1.5 (56 observations) or TNOCB ≥ 1.5 (871 observations). The node with 56 observations is declared a terminal node because of $56 - 0$ majority in favour of "yes". The node with the 871 observations which consist of 263 under-five mortality and 608 not experienced mortality, is then split by whether WI = rich (208 observations) or WI =middle, poor (663 observations). The node with 208 is then split by whether ETH = other (26 observations) or ETH = akan, ewe, ga/adagme, mole-dagbani (182 observations). The node with 26 observations is declared a terminal node for "yes" because of the $26-0$ majority in favour of "yes". The node with 182 observations, which consists of 63 experienced mortality and 119 not experienced mortality, is split by whether PBI < 17.5 (8 observations) or PBI ≥ 17.5 (174 observations). The node with 174 observations which consist of 56 experienced mortality and 118 not experienced mortality, is split by whether DROR = AR,ER,GA,NR,UW, UE (80 observations) or DROR = VR, WR, BA, CR (94 observations). The node with 80 observations is then split by whether AORAB < 17.5(9 observations) or AORAB$\geq$ 17.5 (71 observations). The node with 9 observations is declared terminal node because of $8 - 1$ majority in favour of "yes", and the node with 71 observations, which consist of 26 experienced mortality and 45 not experienced mortality is then split by whether PBI ≥ 67.5 (13 observations) or PBI < 67.5 (58 observations). The node with 13 respondents is declared a terminal node because of $9 - 4$ majority in favour of "yes" whilst the node with 58 observations is also declared a terminal node because of $17 - 41$ majority in favour of "no". The rest of the results are shown in the Figure 4.4.

Estimating the misclassification rate

From Figure 4.4 and Table 4.6, out of the 927 subjects in the data, the classification tree misclassifies 154 of the observations who has experienced under-five

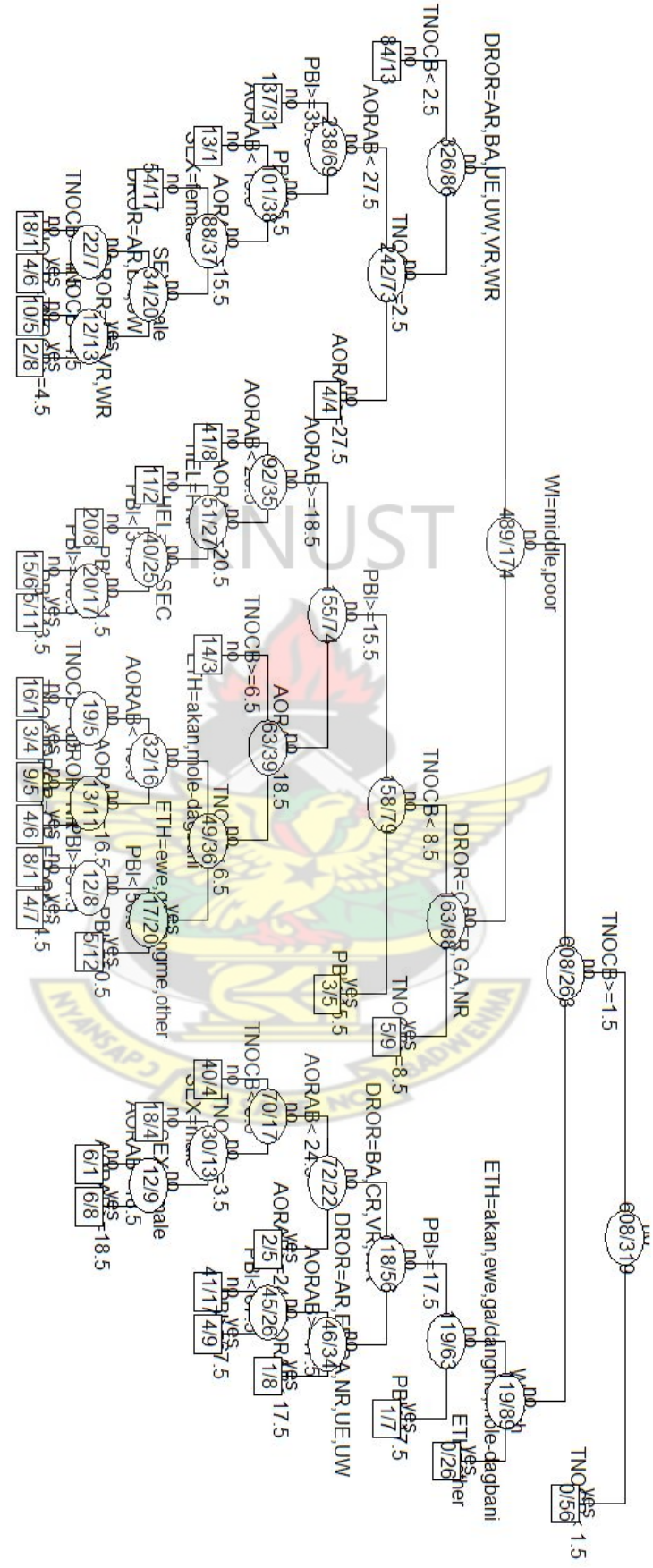


Figure 4.4: *Classification tree for under-five mortality in the rural DHS data*

mortality (yes) as not experienced it before (no), whereas, out 608 who has not experienced mortality before (no), 37 are misclassified as having experienced mortality before (yes) . This gives the resubstitution estimate as

$$R^{re}(T) = \frac{37 + 154}{927} = 0.206 \tag{4.2}$$

Table 4.6: *Misclassification rate for rural*

	Predicted	
Mortality	**Yes**	**No**
Yes	165	154
No	35	571

Pruning the Tree

Table 4.7: *Cross Parameter table for urban*

CP	**nsplit**	**rel error**	**xerror**$(R^{cv/10})$	**xstd**(T_k)
0.1755486	0	1.00000	1.00000	0.045344
0.0407524	1	0.82445	0.82445	0.043026
0.0188088	3	0.74295	0.74295	0.041636
0.0109718	4	0.72414	0.74608	0.041693
0.0094044	7	0.68652	0.74303	0.041087
0.0062696	8	0.67712	0.80564	0.042724
0.0054859	11	0.65831	0.81818	0.042927
0.0047022	16	0.62696	0.84639	0.043364
0.0031348	20	0.60815	0.85266	0.043458
0.0031000	23	0.59875	0.87774	0.043823
0.0020899	31	0.57367	0.90282	0.044169
0.0015674	34	0.56740	0.91223	0.044294
0.0010449	40	0.55799	0.94671	0.044731
0.0007837	43	0.55486	0.95925	0.044882
0.0000000	47	0.55172	0.98433	0.045171

Table 4.7 provides a brief summary of the overall fit of the model for rural data. The relative error (rel error) continues to decrease as the tree becomes more complex, it then begins to rise to some point. This is shown in Figure 4.6. From Table 4.7, the $1 - SE$ rule yields a minimum of $CVerror + SE = 0.74303 + 0.041087 = 0.784117$, which leads to the choice of a classification tree with 7

splits (8 terminal nodes) based upon the cross-validation. The corresponding pruned classification tree is displayed in Figure 4.5.

Variable of importance

According to the discriminatory power in the CART analysis, total number of children born (root node) emerged as the strongest discriminating factor for under-five mortality, followed preceding birth interval (79.97%), Ethnicity (60.00%), Age of respondents at birth (37.88%), wealth index (30.77%), Highest educational level (19.33%) and sex (4.80%) as the least factor that contributes to under-five mortality in Ghana. (Table 4.8).

Table 4.8: *Ranking the predictors of inequalities in under-five mortality in the rural area*

Independent variable	**Power(%)**
Total number of children born	100.00
Preceding birth interval	79.97
Ethnicity	60.00
Age of respondents at first birth	37.88
Wealth index	30.77
Region of Residence	29.00
Highest educational level	19.33
Sex of child	4.80

4.4.4 Analysis on the overall data

The data consist of 1,295 observations. Out of this, 479 experienced under-five mortality and 816 has not experience under-five mortality. The classification tree is displayed in Appendix B where we used the entropy measure as the impurity function for splitting. There are 70 splits and 71 terminal nodes in this tree

The root node (TNOCB) with 1295 observations is split according to whether TNOCB < 1.5 (81 observations) or TNOCB ≥ 1.5 (1214 observations). The node with 81 observations is declared a terminal node because of $81 - 0$ majority in favour of "yes". The node with the 1214 observations which consist of 398

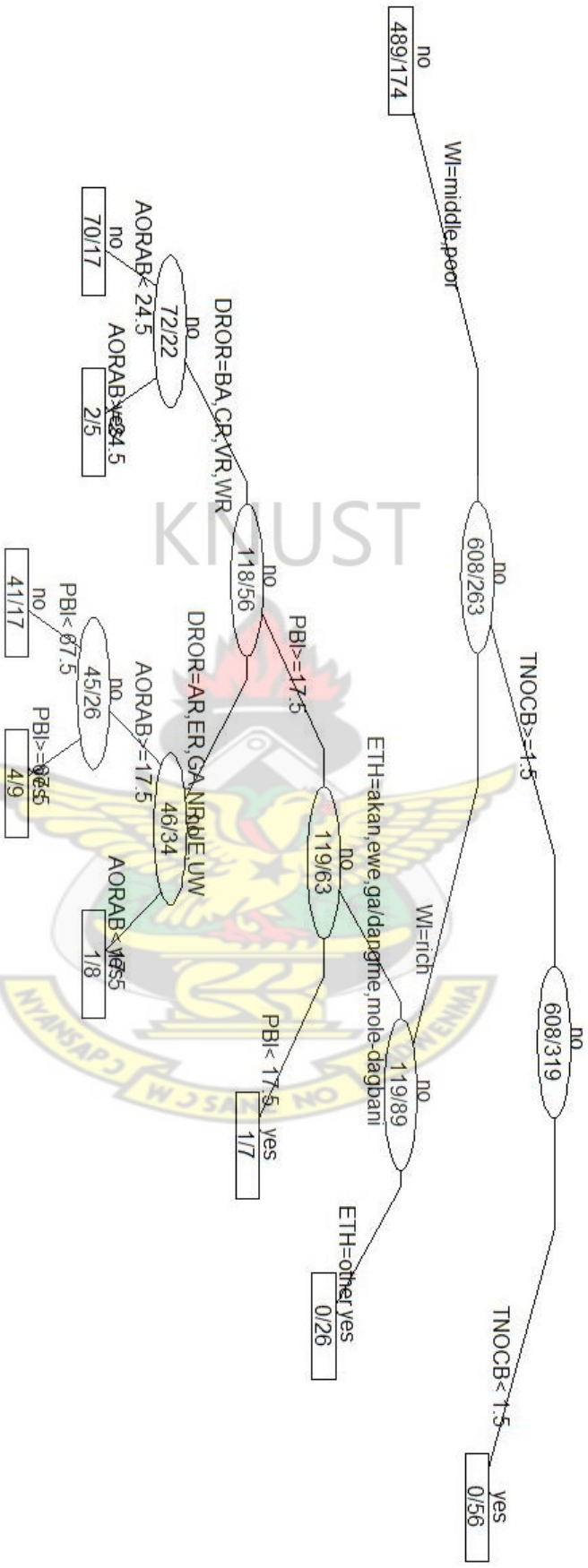


Figure 4.5: *A pruned classification tree with 8 splits and 9 terminal nodes*

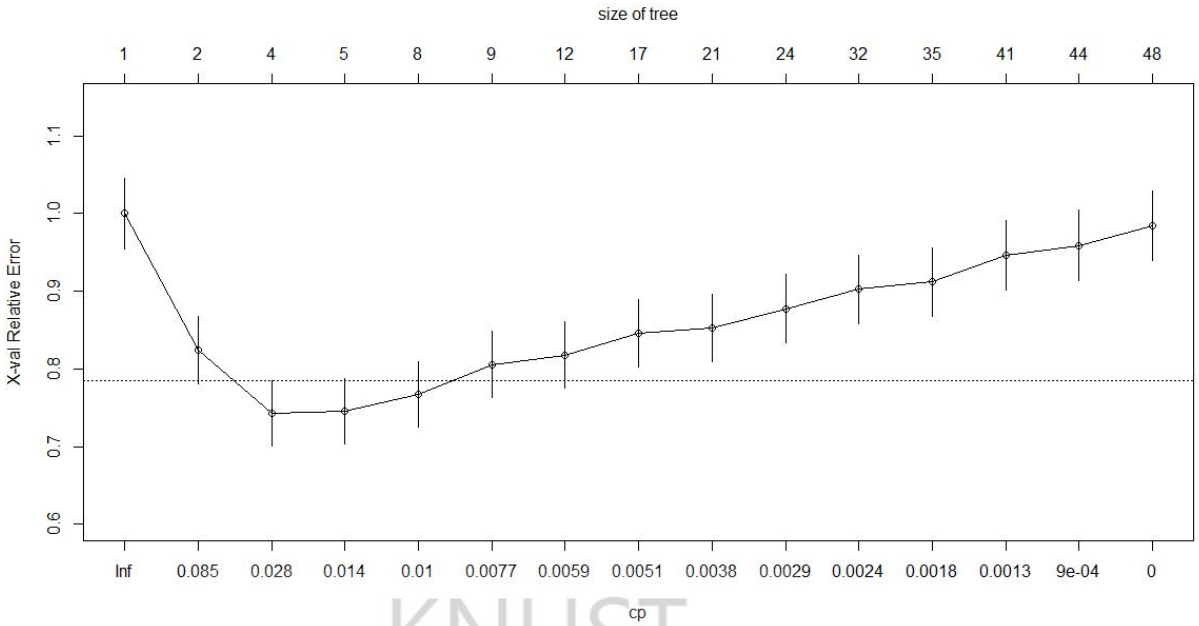


Figure 4.6: *A plot of CP against relative error showing different size of trees for rural*

under-five mortality and 816 not experienced mortality, is then split by whether DROR = GA, NR, UW, WR (443 observations) or DROR = AR, BA, CR, ER, UE, VR (771 observations). The node with 443 is then split by whether DTOR = UR (80 observations) or DTOR = RU (363 observations). The node with 80 observations is then split by whether PBI >= 26 (48 observations) or PBI < 26 (32 observations). The node with 48 observations is declared a terminal node for "yes" because of $48-0$ majority in favour of "yes". The node with 32 observations is then split by whether DROR = NR, UW, WR (9 observations) or DROR = GA (23 observations). The node with 9 observations is declared a terminal node for "yes" because of $9-0$ majority in favour of "yes" whiles the node with 23 observations is also declared a terminal node for "no" because of $4-19$ majority in favour of "no". The rest of the results are shown in the appendix B.

Estimating the misclassification rate

From appendix B and Table 4.9, out of the 1,295 subjects in the data, the classification tree misclassifies 188 of the observations who has experienced under-five mortality (yes) as not experienced it before (no), whereas, out 816 who has not

experienced mortality before (no), 81 are misclassified as having experienced mortality before (yes) . This gives the resubstitution estimate as

$$R^{re}(T) = \frac{81 + 188}{1295} = 0.208 \tag{4.3}$$

Table 4.9: *Misclassification rate for overall analysis*

	Predicted	
Mortality	**Yes**	**No**
Yes	291	188
No	81	735

Pruning the tree

Table 4.10: *Cross Parameter table for urban*

CP	**nsplit**	**rel error**	**xerror**$(R^{cv/10})$	**xstd**(T_k)
0.1691022965	0	1.0000000	1.0000000	0.03626957
0.0438413361	1	0.8308977	0.8308977	0.03466310
0.0156576200	3	0.7432150	0.7453027	0.03357107
0.0135699374	5	0.7118998	0.7348643	0.03342387
0.0052192067	8	0.6680585	0.7118998	0.03308880
0.0041753653	14	0.6346555	0.7286013	0.03333403
0.0031315240	15	0.6304802	0.7703549	0.03391160
0.0020876827	32	0.5657620	0.8058455	0.03436412
0.0008350731	46	0.5219207	0.8956159	0.03536040
0.0006958942	60	0.5010438	0.9144050	0.03554324
0.0002982404	63	0.4989562	0.9311065	0.03569857
0.0000000000	70	0.4968685	0.9331942	0.03571752

Table 4.10 provides a brief summary of the overall fit of the model for the entire data. The relative error (rel error) continues to decrease as the tree becomes more complex. This is shown in Figure 4.8. From table 4.10, the $1 - SE$ rule yields a minimum of $CVerror + SE = 0.7118998 + 0.0330880 = 07449886$, which leads to the choice of a classification tree with 8 splits (9 terminal nodes) based upon the cross-validation. The corresponding pruned classification tree is displayed in Figure 4.7.

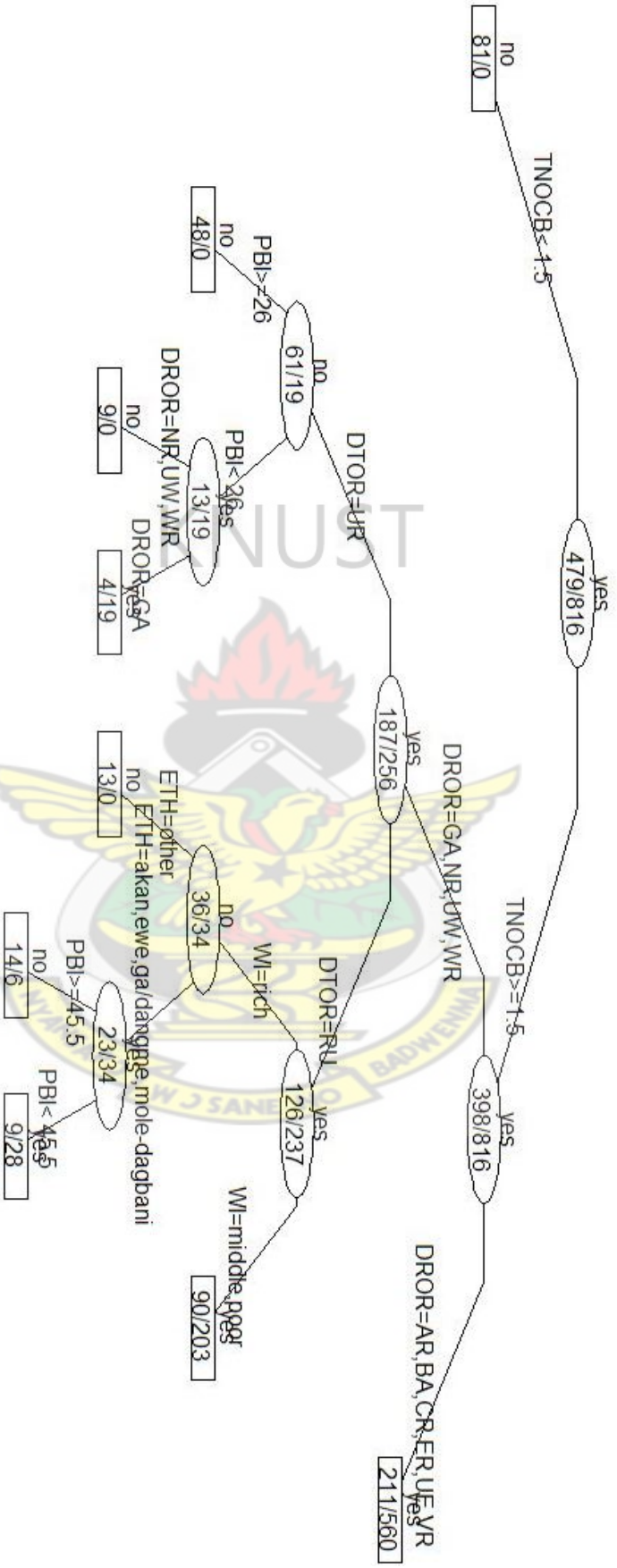


Figure 4.7: *A pruned classification tree for overall analysis*

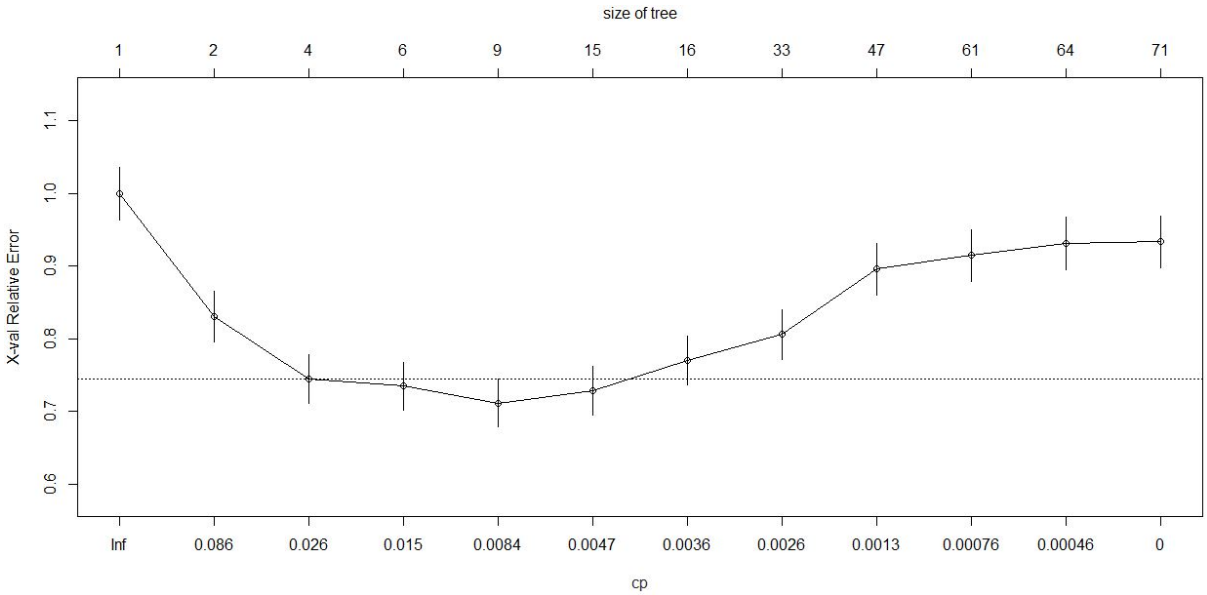


Figure 4.8: *A plot of CP against relative error showing different size of trees for overall analysis*

Variable of importance

According to the overall discriminatory power in the CART analysis, Total number of children ever born which is the root node emerged as the strongest overall discriminating factor for under-five mortality in Ghana, followed by Preceding birth interval (84.60%), Defacto region of residence (57.70%), Ethnicity (38.50%), Age of respondents at birth (30.80%), Defacto type of residence (26.90%),Wealth index (23.10%), Highest educational level (19.23%) and Sex of child (3.8%) as the least contributing factor(Table 4.11).

Table 4.11: *Ranking the predictors of inequalities in under-five mortality in the GDHS data*

Independent variable	**Power(%)**
Total number of children ever born	100.00
Preceding birth interval	84.60
Defacto region of residence	57.70
Ethnicity	38.50
Age of respondents at first birth	30.80
Defacto type of residence	26.90
Wealth index	23.10
Highest educational level	19.23
Sex of child	3.80

4.5 Discussion

This work demonstrates the application of classification tree analysis models which is a non-parametric modeling methodology to explore socio-economic and demographic factors influencing child mortality in Ghana. In this analysis, the classification tree method revealed logical results of the relationships between the outcome of interest(under-five mortality) and the predictor variables. Classification tree analysis helps in determining population segments that need specific attention in relation to the outcome. Segmenting populations supports decision makers in focusing their efforts to specific areas or subgroups. It is important to note that this analysis does not support any claim of superiority of one methodology compared to the other.

In Ghana, most determinants shows a positive contribution socioeconomic inequality in the under-five mortality. This implies the combined effect of the marginal effect of the explanatory variable in under-five mortality and its distribution by economic status is to raise socioeconomic inequality in under-five mortality such that under-five mortality is greater among the poor. This effect may arise either because the particular determinant is more prevalent among people of lower economic status and is associated with a higher under-five mortality risk or because the determinant in question is more prevalent among people of higher economic status and associated with a lower under-five mortality risks.

From the contributions expressed in Table 4.5, it was evident that the highest contributor of under-five mortality in the urban data is age of mother at first birth (AORAB). It's predicting power was 100% even though age of mother at birth did not appear at the terminal node but in the analysis, its contribution to under-five mortality was great. The second predictor was total number of children ever born (TNOCB) (63.7%). It was observed in the classification tree

in Figure 4.2 that mothers whose total birth are less than two contributes positively to under-five mortality ($16 - 0$ majority) than mothers whose total births (TNOCB) are more than one. Preceding birth interval (PBI) which is defined as difference between the current birth and the previous birth was another important predictor of under-five child mortality in Ghana. Mothers whose birth interval are less than 19 months are likely to experience under-five mortality in the urban analysis compared to those whose preceding birth interval is greater than 19 months. This confirms the report by GDHS (2008) which says mothers with first birth and higher order births have an elevated risk of dying. Mothers whose highest educational level (HEL) is up to the primary school have 55.5% chance of experiencing under-five mortality compared to mothers whose educational level is above primary school (23.3%). It was also observed from the urban analysis that children from rich wealth quintiles (WI) have 74.1% chance of not experiencing under-five mortality. But those from poor and middle wealth quintiles have only 40% chance of not experiencing under-five mortality. It was also observed that regions like Northern, Upper East, Upper West, Volta and Western regions of Ghana experience under-five mortality than those in the other regions.

Considering the analysis on the rural data, it was observed that total number of children ever born (TNOCB) is the highest contributor of under-five mortality. From figure 4.4, it is observed that mothers with TNOCB equal to one has greater chance of experiencing under-five mortality than mothers whose TNOCB are greater than one. Preceding birth interval (PBI) was the second highest discriminatory power. It was observed from Figure 4.4 that mothers whose birth interval are less than 17 months and those whose birth interval are more than 68 months experience under-five mortality (87.5%) than those whose birth interval falls within that range. Ethnicity which is a demographic factor shows a positive relation to under-five mortality in the rural data. It was observed that 57.1% of mothers who belongs to ethnic groups like Ewe and Mole-dagbani stands the

chance of experiencing under-five mortality than those in different ethnic groups of Ghana. Age of mother at birth (AORAB) in the rural analysis was also significant (60%) in decomposing inequalities in the under-five mortality. Mothers who age are less than 17 years experience under-five mortality than mothers whose age are above 17 years. Other factors that also contributed positively include highest education level (HEL) and wealth index (WI) with predicting factor of 19.33% and 30.77% respectively.

Considering the overall data, it was clear that TNOCB contribution to under-five mortality was high. Mothers with less than two children experience under-five mortality than those with more than one child. It was also observed that preceding birth interval (PBI) contributed positively to child mortality. Mothers whose birth intervals are less than 16 months stands the chance of experiencing under-five mortality (99%) than those with birth interval above 16 months. From appendix B it was observed that mothers in Northern (87.7%), Greater Accra (69.2%) and Central, Upper East and Upper West (64.7%) regions of Ghana experiences under-five mortality compared to the other regions of Ghana. It was also evident that mothers in the rural areas stands 62.5% chance of experiencing under-five mortality than those in the urban areas of Ghana. Other factors that also contributed positively in the overall analysis are age of mothers at birth (57.8%), ethnicity, wealth index and highest educational levels.

It was observed that sex and whether the child is born by cesarean does not contribute to socio-economic and demographic inequalities in Ghana.

Some differentials were also identified considering the ranking of the predictors of urban and rural. It was identified that Total number of children ever born in the rural areas was the highest contributing factor to under-five mortality in the DHS data compared with that of the urban (63.7%). This was evident

from the overall analysis were total number of children ever born was ranked highest (100%) among all the predictors. Also, there were differentials in age of respondents at birth. This was the highest contributing factor to the inequalities that exist in under-five mortality in the urban area compared with that of the rural area (37.88%) which was ranked fourth in the analysis. Other differentials that were identified include preceding birth interval and Ethnicity. These factors were higher in the rural area 79.97% and 60.00% respectively than in the urban area 59.70% and 30.40% respectively.

Chapter 5

Conclusion and Recommendations

5.1 Conclusion

Among the predictor variables considered in the study to decompose inequality in under-five mortality, it was observed that socio-economic inequalities in under-five mortality favours the better-off (middle and rich quintile) than the poor. These inequalities vary between regions. It was observed that regions such as Northern, Central, Upper East and Upper West contributes positively to under-five mortality than the other regions of Ghana. Other socio-economic factors that affects under-five mortality positively includes mother's educational level (HEL) and Defacto place of residence. The study revealed that child born by caesarean does not contribute to under-five mortality in Ghana.

Demographic factors that was significant to under-five mortality includes Total number of children ever born by mothers, preceding birth interval, age of mother at birth and ethnicity.

It was observed that sex of a child has a negative contribution to under-five mortality in Ghana.

Lastly, there differentials in some socio-economic and demographic factor. these includes total number of children ever born by mothers,age of respondents, preceding birth interval and Ethnicity.

5.2 Recommendations

Since the study used only one method (classification trees), it is recommended that other methodological approach should be used in addition to classification trees to ascertain the best model for this research work.

The analysis by regions highlighted inequalities between regions and wealth-related inequality within regions. A more in-depth analysis determining the location of the most vulnerable sub-groups within regions would help in better reaching the whole population during interventions. It is highly recommended that each intervention should be consistent with the socioeconomic and political context which plays a role in the process to equity illustrated in the conceptual framework proposed by the commission on Social Determinants of Health (CSDH) in the year 2005.

As a future work, it is recommended that determinants of mortality risk among children between the months of $1 - 12$ in Ghana in addition to the determinants of under-five mortality risk can be analyzed. Socio-economic and demographic factors associated with under-five mortality risk might be different than those associated with infant mortality. Therefore explanatory variables other than the ones used in this study can be included to the model.

Lastly,further research should be conducted in subsequent Ghana Demographic and Health Surveys in order to monitor the causes and progress of under-five mortality.

REFERENCES

Austin, P. C. (2008). R and s-plus produced different classification trees for predicting patient mortality. *Journal of Clinical Epidemiology*, 64:1222 –1226.

Breiman L., Friedman, J. H. O. R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. CRC Press. MR0726392.

Hosseinpoor, A. R., van Doorslaer, E., Speybroeck, N., Naghavi, M., Mohammed, K., Majdzadeh, R., Delavar, B., and Jamshidi, H. (2006). Decomposing socioeconomic inequality in infant mortality in iran. *International Journal of Epidemiology*, 35:1211–1219.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer Science Business Media, LLC, 233 New York, NY 10013, USA.

JP, M., I, S., AJ, R., MM, S., G, M., M, L., and AE, K. (2008). socioeconomic inequalities in health in 22 european countries compared the magnitude of inequalities in mortality and self-assessed health among 22 countries in all parts of europe. *pubmed*, 359:e14.

Kajungu, D. K., Selemani, M., Masanja, I., Baraka, A., Njozi, M., Khatib, R., Dodoo, A. N., Binka, F., Macq, J., D'Alessandro, U., and Speybroeck, N. (2012). Using classification tree modelling to investigate drug prescription practices at health facilities in rural tanzania. *Malaria Journal*, 11:1–11.

Kraft, A. D., Nguyen, K.-H., Jimenez-Soto, E., and Hodge, A. (2013). Stagnant neonatal mortality and persistent health inequality in middle-income countries: A case study of the philippines. *Plos One*, 8:1–12.

Malderen, C. V., Ogali, I., Khasakhala, A., Muchiri, S. N., Sparks, C., Oyen, H. V., and Speybroeck, N. (2013a). Decomposing kenyan socio-economic in-

equalities in skilled birth attendance and measles immunization. *International Journal For Equity In Health*, pages 1–13.

Malderen, C. V., Oyen, H. V., and Speybroeck, N. (2013b). Contributing determinants of overall and wealth-related inequality in under-5 mortality in 13 african countries. *Institute of Health and Society (IRSS), Université catholique de Louvain, Brussels, Belgium*, pages 1–10.

Maydana, E., Serral, G., and y Carme Borrell (2009). Socioeconomic inequalities and infant mortality in bolivia. *Pan Am J Public Health*, 25:401–410.

Mosquera, P. A., Hernandez, J., Vega, R., Martinez, J., Labonte, R., Sanders, D., and Sebastian, M. S. (2012). The impact of primary health care in reducing inequalitites in child health outcomes, bogota-colombia: an ecological analysis. *International Journal For Equity In Health*, 11:1–12.

Nagy, K., Reiczigel, J., Harnos, A., Schrott, A., and Kabai, P. (2010). Tree-based methods as an alternative to logistic regression in revealing risk factors of crib-biting in horses. *Journal of Equine Veterinary Science Vol 30, No 1 (2010)*, 30:21–26.

NDPC/GOG (2012). 2010 ghana millennium development goals report. Technical report, UNDP and Government of Ghana/NDPC.

Nkoni (2011). Explaining household socio-economic related child health inequalities using multiple methods in three diverse settings in south africa. *Equity in health*.

P, K., M, H., and L, L. (2006). using classification trees to assess low birth weight outcomes. *Epub*, 38:275–89.

Rahman, K. M. M. and Sarkar, P. (2009). Determinants of infant and child mortality in bangladesh. *Pakistan Journal of Social Sciences*, 6(3):175–180.

S, N., R., H. A., MH, F., B, G., and R, M. (2012). Decomposing socioeconomic inequality in self-rated health in tehran. *J Epidemiol Community Health*, 66:495–500.

SEÇKIN, N. (2009). Determinants of infant mortality in turkey. Master's thesis, MIDDLE EAST TECHNICAL UNIVERSITY.

UN/MDG (2012). The millennium development goals report 2012. Technical report, United Nations.

WAGSTAFF, A. (2000). Socio economic inequalities in child mortality : Comparison across nine developing countries. *Bulletin Of World Health Organization*, 78:19–29.

Whitehead, M. and Dahlgren, G. (1992). Concepts and principles for tackling social inequities in health. Technical report, university of Liverpool.

Zere, E., Kirigia, J. M., Duale, S., and Akazili, J. (2012). Inequities in maternal and child health outcomes and interventions in ghana. *BMC Public Health*, pages 1–10.

APPENDICES

Table A: Chi-Square Tests of association between DTOR and mortality

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	9.290[a]	1	.002
Likelihood Ratio	9.181	1	.002
Fisher's Exact Test			
N of Valid Cases[b]	1295		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 136.12.

b. Computed only for a 2x2 table

5.2.1 appendix B

R-SOFTWARE OUTPUT FOR OVERALL ANALYSIS

n= 1295

node), split, n, loss, yval, (yprob)

∗ denotes terminal node

1) root 1295 479 yes (0.36988417 0.63011583)

2) TNOCB< 1.581 0 no (1.00000000 0.00000000) ∗ 3) TNOCB>= 1.51214398 yes (0.32784185 0.67215815)

6) DROR=GA,NR,UW,WR 443 187 yes (0.42212190 0.57787810)

12) DTOR=UR 80 19 no (0.76250000 0.23750000)

24) PBI>= 2648 0 no (1.00000000 0.00000000) ∗

25) PBI< 26 32 13 yes (0.40625000 0.59375000)

50) DROR=NR,UW,WR 9 0 no (1.00000000 0.00000000) ∗

51) DROR=GA 23 4 yes (0.17391304 0.82608696) ∗

13) DTOR=RU 363 126 yes (0.34710744 0.65289256)

26) WI=rich 70 34 no (0.51428571 0.48571429)

52) ETH=other 13 0 no (1.00000000 0.00000000) ∗

53) ETH=akan,ewe,ga/dangme,mole-dagbani 57 23 yes (0.40350877 0.59649123)

106) PBI>=45.5 20 6 no (0.70000000 0.30000000)

212) ETH=mole-dagbani 10 0 no (1.00000000 0.00000000) *

213) ETH=akan,ga/dangme 10 4 yes (0.40000000 0.60000000) *

107) PBI< 45.5 37 9 yes (0.24324324 0.75675676) *

27) WI=middle,poor 293 90 yes (0.30716724 0.69283276)

54) PBI< 101 283 90 yes (0.31802120 0.68197880)

108) DROR=GA,NR 156 58 yes (0.37179487 0.62820513)

216) PBI< 15.57 2 no (0.71428571 0.28571429) *

217) PBI>=15.5 149 53 yes (0.35570470 0.64429530)

434) AORAB< 18.5 73 32 yes (0.43835616 0.56164384)

868) SEX=male 38 18 no (0.52631579 0.47368421)

1736) PBI >=50.5 14 4 no (0.71428571 0.28571429) *

1737) PBI< 50.5 24 10 yes (0.41666667 0.58333333)

3474) PBI< 34 16 7 no (0.56250000 0.43750000) *

3475) PBI>=34 8 1 yes (0.12500000 0.87500000) *

869) SEX=female 35 12 yes (0.34285714 0.65714286)

1738) ETH=other 13 6 no (0.53846154 0.46153846) *

1739) ETH=akan,ewe,mole-dagbani 22 5 yes (0.22727273 0.77272727) *

435) AORAB>= 18.5 76 21 yes (0.27631579 0.72368421)

870) PBI< 52.5 62 20 yes (0.32258065 0.67741935)

1740) PBI>= 36 22 11 no (0.50000000 0.50000000)

3480) PBI< 42.5 8 1 no (0.87500000 0.12500000) *

3481) PBI>= 42.5 14 4 yes (0.28571429 0.71428571) *

1741) PBI< 36 40 9 yes (0.22500000 0.77500000) *

871) PBI>=52.5 14 1 yes (0.07142857 0.92857143) *

109) DROR=UW,WR 127 32 yes (0.25196850 0.74803150)

218) TNOCB< 5.5 99 29 yes (0.29292929 0.70707071)

436) TNOCB>= 2.5 67 24 yes (0.35820896 0.64179104)

872) AORAB>=16.5 55 22 yes (0.40000000 0.60000000)

1744) PBI>= 23.5 45 20 yes (0.44444444 0.55555556)

3488) PBI< 29.5 10 4 no (0.60000000 0.40000000) *

3489) PBI>=29.5 35 14 yes (0.40000000 0.60000000)

6978) TNOCB>=3.5 19 9 no (0.52631579 0.47368421) *

6979) TNOCB< 3.5 16 4 yes (0.25000000 0.75000000) *

1745) PBI< 23.5 10 2 yes (0.20000000 0.80000000) *

873) AORAB< 16.5 12 2 yes (0.16666667 0.83333333) *

437) TNOCB< 2.5 32 5 yes (0.15625000 0.84375000) *

219) TNOCB>=5.5 28 3 yes (0.10714286 0.89285714) *

55) PBI>= 101 10 0 yes (0.00000000 1.00000000) *

7) DROR=AR,BA,CR,ER,UE,VR 771 211 yes (0.27367056 0.72632944)

14) PBI< 21.5 83 38 yes (0.45783133 0.54216867)

28) ETH=ga/dangme,other 7 1 no (0.85714286 0.14285714) *

29) ETH=akan,ewe,mole-dagbani 76 32 yes (0.42105263 0.57894737)

58) TNOCB< 7.5 68 31 yes (0.45588235 0.54411765)

116) WI=poor,rich 48 23 no (0.52083333 0.47916667)

232) ETH=ewe 8 1 no (0.87500000 0.12500000) *

233) ETH=akan,mole-dagbani 40 18 yes (0.45000000 0.55000000)

466) AORAB>= 21.5 11 3 no (0.72727273 0.27272727) *

467) AORAB< 21.5 29 10 yes (0.34482759 0.65517241)

934) DROR=AR,CR,UE 20 9 yes (0.45000000 0.55000000)

1868) AORAB< 18.5 11 4 no (0.63636364 0.36363636) *

1869) AORAB>= 18.5 9 2 yes (0.22222222 0.77777778) *

935) DROR=BA,ER,VR 9 1 yes (0.11111111 0.88888889) *

117) WI=middle 20 6 yes (0.30000000 0.70000000) *

59) TNOCB>= 7.5 8 1 yes (0.12500000 0.87500000) *

15) PBI>= 21.5 688 173 yes (0.25145349 0.74854651)

30) AORAB>= 13.5 676 173 yes (0.25591716 0.74408284)

60) ETH=other 81 30 yes (0.37037037 0.62962963)

120) WI=middle,rich 48 23 yes (0.47916667 0.52083333)

240) DTOR=RU 24 9 no (0.62500000 0.37500000)

480) WI=rich 11 0 no (1.00000000 0.00000000) *

481) WI=middle 13 4 yes (0.30769231 0.69230769) *

241) DTOR=UR 24 8 yes (0.33333333 0.66666667)

482) TNOCB>= 4.5 7 3 no (0.57142857 0.42857143) *

483) TNOCB< 4.5 17 4 yes (0.23529412 0.76470588) *

121) WI=poor 33 7 yes (0.21212121 0.78787879)

242) HEL=NO 21 7 yes (0.33333333 0.66666667)

484) DROR=AR,BA,VR 8 3 no (0.62500000 0.37500000) *

485) DROR=ER,UE 13 2 yes (0.15384615 0.84615385) *

243) HEL=PRI,SEC 12 0 yes (0.00000000 1.00000000) *

61) ETH=akan,ewe,ga/dangme,mole-dagbani 595 143 yes (0.24033613 0.75966387)

122) PBI< 34.5 171 53 yes (0.30994152 0.69005848)

244) TNOCB>= 3.5 91 37 yes (0.40659341 0.59340659)

488) HEL=NO,PRI 53 26 yes (0.49056604 0.50943396)

976) TNOCB>=5.5 24 9 no (0.62500000 0.37500000)

1952) PBI< 27.5 12 2 no (0.83333333 0.16666667) *

1953) PBI>= 27.5 12 5 yes (0.41666667 0.58333333) *

977) TNOCB< 5.5 29 11 yes (0.37931034 0.62068966)

1954) DROR=AR,ER,UE,VR 21 10 yes (0.47619048 0.52380952)

3908) SEX=male 11 4 no (0.63636364 0.36363636) *

3909) SEX=female 10 3 yes (0.30000000 0.70000000) *

1955) DROR=BA,CR 8 1 yes (0.12500000 0.87500000) *

489) HEL=HI,SEC 38 11 yes (0.28947368 0.71052632)

978) DROR=BA 7 2 no (0.71428571 0.28571429) * 979) DROR=AR,CR,ER,UE,VR

31 6 yes (0.19354839 0.80645161) *

245) TNOCB< 3.5 80 16 yes (0.20000000 0.80000000)

490) AORAB>= 22.5 20 9 yes (0.45000000 0.55000000)

980) HEL=SEC 9 3 no (0.66666667 0.33333333) *

981) HEL=HI,NO,PRI 11 3 yes (0.27272727 0.72727273) *

491) AORAB< 22.5 60 7 yes (0.11666667 0.88333333) *

123) PBI>= 34.5 424 90 yes (0.21226415 0.78773585)

246) PBI>= 36.5 393 88 yes (0.22391858 0.77608142)

492) PBI< 167 386 88 yes (0.22797927 0.77202073)

984) PBI>= 72.5 112 33 yes (0.29464286 0.70535714)

1968) PBI< 106 78 28 yes (0.35897436 0.64102564)

3936) PBI>= 98.5 14 6 no (0.57142857 0.42857143) * 3937) PBI< 98.5 64 20 yes (0.31250000 0.68750000)

7874) DROR=AR,ER 30 14 yes (0.46666667 0.53333333)

15748) PBI>= 87.5 7 2 no (0.71428571 0.28571429) *

15749) PBI< 87.5 23 9 yes (0.39130435 0.60869565)

31498) AORAB< 18.5 10 4 no (0.60000000 0.40000000) *

31499) AORAB>=18.5 13 3 yes (0.23076923 0.76923077) *

7875) DROR=BA,CR,UE,VR 34 6 yes (0.17647059 0.82352941) *

1969) PBI>= 106 34 5 yes (0.14705882 0.85294118) *

985) PBI< 72.5 274 55 yes (0.20072993 0.79927007)

1970) PBI< 58.5 214 48 yes (0.22429907 0.77570093)

3940) ETH=akan 131 34 yes (0.25954198 0.74045802)

7880) AORAB>= 26.5 7 3 no (0.57142857 0.42857143) *

7881) AORAB< 26.5 124 30 yes (0.24193548 0.75806452)

15762) PBI>= 54.5 15 6 yes (0.40000000 0.60000000) *

15763) PBI< 54.5 109 24 yes (0.22018349 0.77981651)

31526) PBI< 38.5 15 6 yes (0.40000000 0.60000000) *

31527) PBI>=38.5 94 18 yes (0.19148936 0.80851064)

63054) AORAB< 23.5 87 18 yes (0.20689655 0.79310345)

126108) WI=poor,rich 65 16 yes (0.24615385 0.75384615)

252216) PBI>= 43.5 43 13 yes (0.30232558 0.69767442)

504432) TNOCB>= 3.5 20 8 yes (0.40000000 0.60000000)

1008864) DROR=AR,ER 9 4 no (0.55555556 0.44444444) *

1008865) DROR=BA,CR 11 3 yes (0.27272727 0.72727273) *

504433) TNOCB< 3.5 23 5 yes (0.21739130 0.78260870) *

252217) PBI< 43.5 22 3 yes (0.13636364 0.86363636) *

126109) WI=middle 22 2 yes (0.09090909 0.90909091) *

63055) AORAB>= 23.5 7 0 yes (0.00000000 1.00000000) *

3941) ETH=ewe,ga/dangme,mole-dagbani 83 14 yes (0.16867470 0.83132530) *

1971) PBI>= 58.5 60 7 yes (0.11666667 0.88333333) *

493) PBI>= 167 7 0 yes (0.00000000 1.00000000) *

247) PBI< 36.5 31 2 yes (0.06451613 0.93548387) *

31) AORAB< 13.5 12 0 yes (0.00000000 1.00000000) *