# DEVELOPING A DATA MINING CLASSIFICATION MODEL TO PREDICT THE ACADEMIC PERFORMANCE OF STUDENTS IN PUBLIC BASIC SCHOOLS IN GHANA USING SOCIO-ECONOMIC VARIABLES. A CASE STUDY OF SELECTED PUBLIC BASIC SCHOOLS IN THE ABLEKUMA WEST CONSTITUENCY.



**BY**

**ALEXANDER SARPONG (BSC Information Technology)**

**A Thesis submitted to the Department of Computer Science,**
**Kwame Nkrumah University of Science and Technology**
**In partial fulfillment of the requirements for the degree of**

**MASTER OF SCIENCE**
**Faculty of Physical Sciences,**
**College of Science**

**November, 2014**

## DECLARATION

I Alexander Sarpong hereby declare that this thesis, with the exception of quotations and references contained in published works which have been duly identified and acknowledged, is entirely my own original work, and it has not been submitted for award of any certificate, diploma or degree in this or any other University or institution.

Sign…… Date……

## DEDICATION

This work is dedicated to the God Almighty with due reverence to my Lord Jesus Christ and the mighty Holy Spirit. My lovely wife Charlotte O. Sarpong, my mother Mrs. Rose Osei and my mother-in-law, Mrs. Vida Nkansah. I also dedicate it to my two wonderful boys: Melville and Medwin for their moral support and daily encouragement, my lecturers and colleagues who greatly supported me to come up with this piece of study.



ALEXANDER SARPONG (PG8310212)

| ……... | …… | …….. |
|---|---|---|
| Student Name &ID | Signature | Date |

Certified by:

| …… | …… | ..…….... |
|---|---|---|
| Supervisor(s) Name | Signature | Date |

Certified by:

| ……..... | …… | ….…… |
|---|---|---|
| Head of Dept. Name | Signature | Date |

**ACKNOWLEDGEMENTS**

The Almighty God receives the highest appreciation and acknowledgement for providing me with sufficient energy, time and wisdom through my Lord Jesus Christ to write up this dissertation. I give you the glory and honour for bringing me this far. Thank you, God. I would like to also express my deepest gratitude to my supervisor, Mr. Frimpong Twum, for his inspirational guidance throughout this project; to the head of the computer science department, Dr. Michael Asante for his guidance and support; to Mr Kwaku Pabbi and all the other lecturers for their wise counsel. I am also very grateful to my family for bringing meaning to my experience at KNUST, I fully acknowledge their patience, tolerance and understanding: Mrs. Charlotte O. Sarpong (wife), Melville and Medwin Sarpong (Sons). These people had to put up with seeming lack of due attention during the duration of my course. I heartly thank them. Special thanks go to all respondents for being flexible in giving me enough information to make this work a success. I owe a debt of gratitude to them all. To the headteachers and teachers of those schools, I say a big thank you. On a personal level, I would like to thank all members of my extended family. I also wish to express my utmost appreciation to my course-mates, for the words of encouragement, which enabled me to carry on during difficult moments of the course. I do recognize your moral and intellectual support. Thank you very much. I also acknowledge all those who contributed in various ways to the study. Thank you Lord for making this dream come true.

## ABSTRACT

The use of educational related data is often beneficial in data mining applications and it has proven to be useful to both decision-making processes and the promotion of social goals. Most developing nations are concentrating on ways to use Information Systems as platforms to champion their national development agenda in all areas of their economy, including education. Despite the high percentage of trained teachers in the public basic schools, results from the West African Examinations Council (WAEC) indicates that public basic schools fare poorer in the Basic Education Certificate Examination (BECE) than their private basic schools counterparts. This thesis focuses on using socio-economic variables to develop a data mining classification model that can be used to identify students from poor socio-economic backgrounds and help improve their performance before writing the Basic Education Certificate Examination. The population for this study comprised of 800 junior high school students whilst a convenient sample of 200 students are used for this study. The CRISP-DM (Cross-Industry Standard Process for Data Mining) is used as a solid framework for guiding the project because of its non-proprietary and neutral background. Three popular algorithms are discussed and the C4.5 algorithm is chosen as the preferred algorithm because of its level of accuracy on unseen data. These algorithms are Naïve Bayes, ID3 and C4.5. The C4.5 algorithm is used to analyze the training set and build a classifier that is used to correctly classify both the training and test examples. A standard machine learning technique is used to analyze the training data and test the accuracy of the hypothesis in predicting the categorization of unseen examples with the test data. This testing process is further boosted by deploying the use of the ROC graph to aid in visualization. This graph is used to present a graphical presentation of the relationship between sensitivity and specificity and to decide on the models optimality through the determination of the best threshold for the classifier. Sensitivity, Specificity and Accuracy are used to measure the correctness of the model by calculating for the True and False Positives and Negatives (Type I and Type II error). The model achieved an accuracy rate of 74%, a recall (R) of 73%, specificity of 75% and a precision of 80%. This study has demonstrated the practicality and feasibility of classifying student academic performance based on the selected socio-economic variables.

## TABLE OF CONTENTS

## LIST OF TABLES

**Table** **Page**

## LIST OF FIGURES

**Figure** **Page**

## LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| **BECE** | – | Basic education certificate education |
| **CRISP-DM** | – | Cross industry standard process for data mining |
| **DM** | – | Data mining |
| **EDM** | – | Educational data mining |
| **FN** | – | False negative |
| **FP** | – | False positive |
| **FS** | – | Family size |
| **GES** | – | Ghana education service |
| **ID3** | – | Iterative dichotomiser 3 |
| **KNN** | – | K-nearest neighbors |
| **MI** | – | Monthly income |
| **MOE** | – | Ministry of education |
| **ODM** | – | Organizational data mining |
| **PEL** | – | Parent educational level |
| **PES** | – | Parent employment status |
| **PMS** | – | Parent marital status |
| **ROC** | – | Receiver operating characteristics |
| **SHS** | – | Senior high school |
| **TP** | – | True positive |
| **TN** | – | True negative |
| **WAEC** | – | West African examinations council |

# CHAPTER ONE

# INTRODUCTION

## 1.1 Background to the study

There is tremendous pressure on educational institutions to provide up-to-date information on institutional effectiveness (Romero & Ventura, 2010). Institutions are also increasingly held accountable for student success (Campbell & Oblinger, 2007). One very important response to this pressure is finding new ways to apply analytical and data mining methods to educationally related data. Data mining techniques provide a promising tool to analyze these factors because they are used to discover hidden patterns and relationships that may be helpful in decision making. Every child has the capability to be successful in school and in life yet far too many children fail to meet their potential. Being able to classify students based on the socio-economic challenges they face is an important step in child development in any educational system. Teachers, school administrators and parents have always wanted to know how their students are doing in the classroom. Recently, interest in tracking student learning has grown dramatically due to increased emphasis on accountability in educational settings. In Ghana, performance remain a problem especially at the public basic schools level where poor performance deprive the country of the much needed educated youth prepared for work and for further education and training. A lot of money is spent on education but this does not reflect positively in the kind of students that come out of the public basic schools. Ghana is a developing country therefore financial resources are scarce to come by due to the fact that other sectors of the economy are craving to have a share of the national cake. There should therefore be significant improvement in the quality of students that come out of our pubic basic schools each year. Socio-economic factors have always played a significant role in the success of children in public funded schools. A person's education is closely linked to their life chances, income, and well-being (Battle and Lewis, 2002). It is therefore important to have a clear understanding of what benefits or hinders one's educational attainment. (Graetz, 1995) carried out a study on social and economic status in educational research and policy formulation and found out that, social and economic challenges remain one of the major sources of educational inequality and add that one's educational success depends very strongly on the socio-economic status of one's parents. (Considine and Zappala, 2002) agree with (Graetz, 1995) in their study on the influence of social and economic challenges in the academic evaluation of school students in Australia found that families where the parents are high on the educational, economic and social ladder foster a higher level of achievement in their children. They also found that these parents

provide a higher degree of psychological support for their children through processes that encourage the development of the necessary skills to be successful at school. There are other topical areas that are most commonly linked to a student's academic performance. These include family size, parent's educational background, parent's employment status, parent's marital status, family income, teacher's punctuality, availability of learning resources, teacher motivation, size of classroom etc. all these factors are important influences on student performance and have been shown to affect examination grades. In addition environmental factors such as school size, neighborhood, and relationships between teachers and students also influence examination grades (Crosnoe, et al., 2004). Research has also found that socio-economic status, parental involvement, and family size are particularly important family factors (Majorbanks, 1996). According to the Ghana Education Service (GES) basic education policy-framework, basic school education should provide the opportunity for students to discover their interests, abilities, aptitudes and other potentials. It should introduce students to basic scientific and technical knowledge and skills and prepare them for further academic work and acquisition of technical and vocational skills at the senior high/technical school level. Therefore, a major priority now is improving quality and student learning outcomes. One major way this can be achieved is by analyzing the effect of socio-economic challenges on students' academic performance and using the patterns identified to help stakeholders formulate policies that will enable students in the public basic schools compete favorably with their counterparts in the private basic schools.

### 1.1.1 Why Data mining

Data mining, also known as Knowledge Discovery in Database, refers to extracting or "mining" knowledge from large amounts of data. Organizations have a challenge of sifting through all of that information, and need solutions to do so. Data mining can assist organizations with uncovering useful information in order to guide decision-making Kiron et al., (2012). Data mining is a series of tools and techniques for uncovering hidden patterns and relationships among data (Dunham, 2003). It can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students Alaa el-Halees, (2009). While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm and Nearest Neighbor method are used for knowledge discovery from databases. Data mining has been applied in a variety of industries, government, military, retail,

and banking but has received much attention in educational contexts (Ranjan & Malik, 2007). Mining in educational environment is referred to as Educational Data Mining (EDM). EDM is a field of study that analyzes and applies data mining to solve educationally-related problems. Applying data mining this way can help researchers and practitioners discover new ways to uncover patterns and trends within large amounts of educational data. The process of data mining uses machine learning, statistics, and visualization techniques to discover and present knowledge in a form that is easily comprehensible. One major objective of the data mining process is prediction. That is, to predict unknown or future values of the attributes of interest using other attributes in the databases, while describing the data in a manner understandable and interpretable to humans. Classification is one of the most useful predictive data mining techniques used with educational learning because it maps data into predefined groups of classes.

**1.2 Supervised learning (classification)**

Supervised methods are methods that attempt to discover the relationship between input attributes (independent variables) and a target attribute (dependent variable). We have a training set of examples with labels, and a test set of examples with unknown labels. The whole point is to make predictions for the test examples, (Elkan, 2012). Decision tree algorithms represent supervised learning, and as such require preclassified target variables. A training data set must be supplied which provides the algorithm with the values of the target variable, (Larose, 2005). The relationship discovered is represented in a structure referred to as a model, (Maimon & Rokach, 2005). In data mining, models describe and explain phenomena, which are hidden in the dataset and can be used for predicting the value of the target attribute knowing the values of the input attributes, (Maimon & Rokach, 2005). This training data set should be rich and varied, providing the algorithm with a healthy cross section of the types of records for which classification may be needed in the future. Larose (2005). The target attribute classes must be discrete. One cannot apply decision tree analysis to a continuous target variable. Rather, the target variable must take on values that are clearly demarcated as either belonging to a particular class or not belonging, Larose (2005). According to Elkan (2012), a common rule of thumb is to use 70% of the database for training and 30% for testing. Every training algorithm looks for patterns in the training data, i.e. correlations between the features and the class. However in research, there is the need to also measure the performance achieved by a learning algorithm and to do this, we use a test set consisting of examples with known labels. We train the classifier on the training set, apply it to the test set, and then

measure performance by comparing the predicted labels with the true labels, Elkan (2012). Only accuracy measured on an independent test set is a fair estimate of accuracy on the whole population.

### 1.3 Statement of the problem

The Basic Education Certificate Examination is administered by the West African Examinations Council in Ghana. To qualify for the examination, a student should have completed three years of pre-school education, six years of primary education and three years of Junior high school. Since the inception of this examination in 1990 the performance of candidates from the public basic schools has not been encouraging despite the fact that they have a high percentage of trained teachers in Ghana. Literatures from other researchers have suggested that, socio-economic challenges remain one of the factors of poor academic performance. For the purposes of this research, the focus is to develop a data mining classification model to classify student's academic performance based on these socio-economic challenges. This model can then help predict the performance of other children with similar challenges thereby enabling school authorities to come up with polices that will help such students perform better in the Basic Education Certificate Examination. A cursory look at the table below shows that the number of trained teachers in the public primary schools far exceeded those in the private primary schools. "The percentage of trained teachers in public junior high schools in the 2011/2012 academic year is 82.9%, when compared to the 2010/2011 figure of 78.4%. In private junior high schools however, the percentage of trained teachers is only 19.8%, when compared to the 2010/2011 figure of 20.3%". (Statistics, Research, Information Management and Public Relations (SRIMPR, 2012 p. 20)

Table 1.1 percentage of trained teachers in the private and public basic schools.

| Type of education | % of trained teachers | |
|---|---|---|
| | 2010/2011 | 2011/2012 |
| Public | 78.4 | 82.9 |
| Private | 20.3 | 19.8 |

#### 1.3.1 Performance Analysis

The tables below show the performance indicators of public and private basic schools in the Ablekuma west constituency of the Greater Accra region from 2011 to 2013. 10 public basic schools and 10 private basic schools were selected to be the representatives of this analysis. The analysis comprised of five core subjects, English, Social studies, RME, Mathematics, Science and one elective subject (ICT) and the number of aggregate one's (1) attained by students of both public and private schools and the corresponding percentage margins. (WAEC, Accra).

Table 1.2 Percentage margins of students with aggregate one (1) in 2011

| | English Language | % | Social Studies | % | Religious and moral education | % | Mathematics | % | Integrated Science | % | ICT | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **No of students in Private schools with aggregate. 1** | 439 | **90.5** | 217 | **89.7** | 207 | **96.3** | 279 | **87.7** | 329 | **89.2** | 308 | **99.7** |
| **No of students in Public schools with aggregate. 1** | 46 | **9.5** | 25 | **10.3** | 8 | **3.7** | 39 | **12.3** | 40 | **10.8** | 1 | **0.3** |
| **Total** | **485** | 100 | **242** | 100 | **215** | 100 | **318** | 100 | **369** | 100 | **309** | 100 |



Table 1.3 Percentage margins of students with aggregate one (1) in 2012

| | English Language | % | Social Studies | % | Religious and moral education | % | Mathematics | % | Integrated Science | % | ICT | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **No of students in Private schools with aggregate. 1** | 486 | **95.5** | 262 | **89.4** | 176 | **86.7** | 355 | **93.4** | 341 | **96.9** | 235 | **97.9** |
| **No of students in Public schools with aggregate. 1** | 23 | **4.5** | 31 | **10.6** | 27 | **13.3** | 25 | **6.6** | 11 | **3.1** | 5 | **2.1** |
| **Total** | **509** | 100 | **293** | 100 | **203** | 100 | **380** | 100 | **252** | 100 | **240** | 100 |

Table 1.4 Percentage margins of students with aggregate one (1) in 2013

| | English Language | % | Social Studies | % | Religious and moral education | % | Mathematics | % | Integrated Science | % | ICT | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **No of students in Private schools with aggregate. 1** | 338 | **88.3** | 283 | **93.1** | 251 | **96.3** | 278 | **95.4** | 308 | **83.9** | 183 | **99.5** |
| **No of students in Public schools with aggregate. 1** | 45 | **11.7** | 21 | **6.9** | 12 | **3.7** | 64 | **4.6** | 59 | **16.1** | 1 | **0.5** |
| **Total** | **383** | 100 | **304** | 100 | **263** | 100 | **342** | 100 | **367** | 100 | **184** | 100 |

From the statistics provided, it is quite evident that, the performance of the public basic schools is very discouraging compared to the private basic schools.

## 1.4 Justification

It is anticipated that the findings and recommendations of this study would go a long way in generating the much needed information that would help public basic schools reform by identifying the weaknesses that exist in their schools and work assiduously to resolve them. The accuracy in performance evaluation has the benefit of making replication much more feasible for schools because once a construct of educational interest has been empirically defined in data; it can be transferred to new data sets. By collecting and analyzing scientific data about these important topics in education, this research can establish the best practices that teachers, students, parents, counselors, administrators, and government can use to improve learning outcomes and boost performance. It will also enable public basic schools reduce the complexity of computation involved with the student academic evaluation process while maintaining high prediction accuracy.

This study will enable educational institutions collect students' academic performance classification data for future educational research purposes and increase opportunities to examine and understand factors that can positively or negatively affect a schools' progress. By making available to teachers information previously obtained through hard copies, this study can increase teacher's familiarity with students and help inform classroom practice. It will also enable school authorities to streamline the educational process by offering proper counseling to students and parents thereby ensuring that parents who have gone through the counseling process set their priorities such that the socio-economic needs of their children will be their topmost priority. Through this study, the Ghana education service can monitor and compare progress in implementing education plans among public basic schools and enable school authorities manage classroom processes according to Ghana Education Service procedures. Lastly, the study is expected to add to the existing body of knowledge and act as a stepping-stone for later researchers in similar studies. It would also help future researchers who have the interest of improving the teaching and learning processes in public basic schools across Ghana.

## 1.5 Project Aim

This study seeks to use data mining to develop a classification model based on socio-economic variables to predict the academic performance of students in public basic schools.

## 1.6 Specific Objectives

The interdependency of the following research objectives will be used in developing the academic performance classification model.

- To use size of family as a socio-economic variable to classify the academic performance of students in public basic schools.
- To use parents' education level as a socio-economic variable to classify the academic performance of students in public basic schools.
- To use parents' employment status as a socio-economic variable to classify the academic performance of students in public basic schools.
- To use parents' marital status as a socio-economic variable to classify the academic performance of students in public basic schools.
- To use family monthly income as a socio-economic variable to classify the academic performance of students in public basic schools.



### 1.7 Research Questions

For the purpose of this study, the following research questions have been formulated to help develop an accurate academic performance classification model.

- Will the model be able to correctly classify student's academic performance using their family size as a socio-economic variable?
- Will the model be able to correctly classify student's academic performance using their parents' education level as a socio-economic variable?
- Will the model be able to correctly classify student's academic performance using their parents' employment status as a socio-economic variable?
- Will the model be able to correctly classify student's academic performance using their parents' marital status as a socio-economic variable?
- Will the model be able to correctly classify student's academic performance using their family monthly income as a socio-economic variable?

### 1.8 Project Scope

Geographically, the study covered the Ablekuma west constituency of the Greater Accra region, Ghana. The boundary of this study focused on certain aspects of socio-economic factors which among others included family size, parent's education level, parent's employment status, parent's marital status and family income. These factors were chosen for the study because a considerable amount of research has been conducted on them by other researchers from other continents and I intend to do same using the Ghanaian environment. The respondents in the study are students of the selected public basic

schools. I did not have outmost control over the participants or the teachers who administered the questionnaire to the students neither did I have control over student's absence when the questionnaire were administered. This limitation might reduce the sample of the study. It was not possible to include final year students since they are on the verge of writing their final exams and will not find the outcome of this study beneficial. The data from the questionnaire was obtained at a certain specific point in time; a parents socio-economic standing may have been improved due to several factors. The researcher therefore has no control over such issues.

### 1.8.1 Assumptions

The major assumption made in this research is that respondents answered the questionnaires truthfully and that they fully understood what the questionnaire required of them. The participants of this study are volunteers and may withdraw from the study at any time and with no ramifications. I also assume that the questionnaire accurately captured the information needed to undertake this study successfully. This research represents a sample of public basic school students from the Ablekuma west constituency of the Greater Accra region; I assume that this sample is representative of the population I wish to make inferences to.

## 1.9 Beneficiaries

- The direct beneficiaries of this study are students, teachers and school administrators who can predict the performance trends of students from poor socio-economic backgrounds and come up with policies to address those challenges.
- This study will enable the Ghana Education Service to come up with policies that will address the challenges of students from poor socio-economic families in order for them to compete favorably with their counterparts in the private basic schools.
- The Ministry of Education (MOE) will benefit from this study by adopting the application of technology in the student academic performance evaluation process.
- This study will make it expedient for the Ministry of education to easily and efficiently classify student performance trends in public basic schools.

## 1.10 Structure of the Thesis

Chapter 1 looks at the background to the study, the problem statement, motivation, project aim and objectives, research questions, scope and beneficiaries. Chapter 2 reviews related works on the test variables, theoretical framework for the academic performance classification model and research hypotheses in data mining. Chapter 3 presents the methodology used in this study and it looks at the

study population and sample size, the sampling technique, research instruments, research design, data collection, demographic relationships and study variables, reliability and validity, measurement procedures, likert scale, data interpretation for training and test set, CRISP-DM (Cross-Industry Standard Process for Data Mining) and ethical Consideration. Chapter 4 focuses on model development and evaluation processes this includes the concept of information theory, entropy and information gain, data modeling (classification ), decision rule extraction, confidence and support, decision rule coverage and measures of classification success. Chapter 5 looks at, summary, conclusions and recommendations.

# CHAPTER TWO

# REVIEW OF RELATED LITERATURE

## 2.0 Introduction

In order to more easily discuss the current state of our education system in relation to educational data mining, it is useful to first look at the history that has brought educational research and data mining technologies together. Educational data mining (EDM) is an emerging discipline that focuses on applying data mining tools and techniques to educationally related data (Baker &Yacef, 2009). Researchers within EDM focus on topics ranging from using data mining to improve institutional effectiveness to applying data mining in improving student learning processes. This chapter reviews the background of educational data mining, application of data mining in educational research, effective use of school data and the socio-economic factors of poor student performance. In this way, it is possible to highlight the important contributions and provide a starting point for my research.

## 2.1 Background to educational data mining

Applying data mining in educational research is a recent research area; there have been lots of research conducted in this area because of its potential to educational institutions. Romero and Ventura, conducted a survey on educational data mining between 1995 and 2005. They came to the conclusion that educational data mining is a promising area of research that has specific requirements not found in other research areas. Educational data mining (EDM) has several definitions, Campbell and Oblinger (2007) defined academic analytics as the use of statistical techniques and data mining in ways that will help faculty and advisors become more proactive in identifying at-risk students and responding accordingly. Academic analytics is considered as a sub-field of EDM and it focuses on processes that occur at the primary, basic, secondary and university level. One of the biggest unresolved challenges facing the basic education system in Ghana is how to make significant improvements in the learning achievements of students from all backgrounds at all levels of schooling (Donge, 2003). Baker and Yacef (2009) also defined EDM as an emerging discipline that is concerned with developing methods for exploring the unique types of data that come from educational institutions, and using those methods to better understand students, and the settings which they learn in (Baker &Yacef, 2009). Data mining is not mentioned in this definition and this leads other researchers to explore and come up with other analytical methods that can be applied to EDM. Educational data mining is therefore a broader term

that looks at any type of data used in educational institutions. The boundaries of EDM include areas that affect students life's directly. EDM also looks at other processes like student admissions, academic performance and teacher effectiveness. According to Anamuah-Mensah, (1997).'teacher education plays a crucial role in empowering a group of people to assist the greater majority of individuals to adapt to the rapidly changing social, economic and cultural environment to ensure the development of human capital required for the economic and social growth of societies'. Data mining techniques such as multivariate statistics, association rule mining and classification are also key techniques applied to educationally related data (Calders & Pechenizkiy, 2012). These data mining tasks are simply methods for conducting exploratory analysis that can be used for effective institutional learning. These tasks can be used for modeling individual student differences and provide a way to respond to those differences thereby improving student learning (Corbett, 2001). Guan et al. (2002) discussed how important it is to have meaningful information available for decision-makers within educational institutions. It is always difficult to get the information that decision makers need to quickly and efficiently make informed decisions. EDM can also adopt ideas from Organizational Data Mining (ODM) which focuses on assisting organizations with sustaining competitive advantage (Nemati & Barko, 2004). The main difference between DM and ODM is that ODM relies on organizational theory as a reference discipline (Nemati & Barko, 2004). Organizations and institutions that process their data into useful information gain huge benefits such as enhanced decision-making, increased competitiveness, and potential financial gains (Nemati & Barko, 2004). EDM can therefore draw upon some of the strengths of organizational theory. Qualitative techniques such as document analysis and interviews are used to support research work in EDM but the dominant research paradigm is quantitative, where results come in the form of predictions, clusters, classifications or associations. Data mining employs statistics, machine learning, and artificial intelligence techniques and this is why research conducted in EDM mainly focuses on quantitative analyses.

### 2.2 Application of data mining in educational research

There are many tasks that can be accomplished in educational institutions by applying data mining techniques. Baker (2009, 2010) suggested four key areas of application for EDM: improving student models, improving domain models, studying the pedagogical support provided by learning software and scientific research into learning and learners. According to Nsiah-Gyabaah (2009), 'there has never

really been any argument over the link between education and development because education helps to build national capacity to apply science and technology to social and economic problems'. Education is a fundamental human right and it is necessary for socio-economic development of society. It is a means to the fulfilment of an individual and the transfer of values from one generation to the next. Castro et al (2007) suggested the following EDM tasks: applications' dealing with the assessment of the student's learning performance, applications that provide course adaptation and learning recommendations based on the student's learning behavior and applications that involve feedback to both teacher and students in e-learning courses. Several studies have analyzed student performance in educational institutions. These studies have mainly focused on classifying students into two categories – either pass or fail for a given course. In a representative study, Kotsiantis et al. (2003) used multiple machine learning techniques to classify university students into dropouts and non-dropouts. The study shows that it is possible to classify the dropout-prone students by using only students' demographic data. Minaei-Bidgoli et al. (2003) conducted a similar study where they tried to predict the final test grades for students enrolled in a web-based course. In this study, three different classifications for the students' results were used: dividing results into two classes (pass and fail), three classes (high, middle and low), or into 9 classes, according to their grade. Several learning algorithms were compared: decision trees, neural networks, naïve Bayes, logistic regression, support vector machines, and k-Nearest Neighbors with feature weights adjusted by a genetic algorithm. The most applied Data Mining tasks are classification, association rule mining, regression and clustering because they yield new insight by uncovering patterns and relationships that they had not previously noticed or considered. The most used DM techniques by several researchers are bayesian networks which model uncertainty by explicitly representing the conditional dependencies among various components, thus providing a graphical visualization of the dependency relationships among the components, neural networks which is a powerful technique for representing complex relationships between inputs and outputs and decision trees which play well with other modeling approaches such as regression and can be used to select inputs or create dummy variables representing interaction effects for regression equations. For example, Neville (1999) explains how to use decision trees to create stratified regression models by selecting different slices of the data population for in-depth regression modeling El-Halees (2008) undertook a case study that used educational data mining to analyze students' learning behavior. The

aim of his study was to show how useful data mining can be used in education to improve students' performance. He used students' data from a database made up of personal and academic records, course records and data from an e-learning system. He now applied data mining techniques to discover many kinds of knowledge such as classification rules using decision tree and association rules. Al-Radaideh et al. (2006) applied classification to help in improving the quality of teaching and learning by evaluating student data to discover the main attributes that may affect student performance in examinations. They also applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5, and the Naive Bayes were used and the results indicated that the Decision Tree model had better prediction than other models. Baradwaj and Pal (2011) also applied classification as a data mining technique to evaluate the performance of students. They used the decision tree method to conduct the classification. The main aim of their research was to extract knowledge that best describes the students' performance in examinations. They selected 300 students from 5 different degree colleges. Using Bayesian classification method on 17 attribute, it was found that the factors like living location, medium of teaching, mother's qualification, family annual income and student's family status were highly correlated with student academic performance. Their study helped in identifying students who needed attention and enabled the teachers to provide the necessary counseling and advice. Quality teachers and quality teaching are some of the most important determinants of a good education. The success of students in education and the progress of the nation will depend on quality teaching which ensures the development of the innate capacities of all students. (GOG, 2002). Shannaq et al. (2010) used classification as data mining tool to predict the numbers of students enrolled in a school by evaluating their academic records to discover the main factors that may affect their loyalty. Chandra and Nandhini (2010) applied association rule based on students' failed courses to identify their failure patterns. Their research was to identify hidden relationship between the failed courses and come up with reasons for the failure so as to help improve the performance of poor performing students. The association rules extracted revealed some hidden patterns of students' failure and this could help policy makers in the education sector make informed policy decisions to help address that challenge. Ayesha et al. (2010) also applied k-means clustering algorithm as a data mining tool to predict students' learning activities that include assignments and final examination.

### 2.2.1 Data mining algorithms

#### 2.2.1.1 ID3

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan in (1986) to generate a decision tree from datasets. ID3 is typically used in machine learning and natural language processing domains. It constructs the decision tree by employing a top-down, greedy search through the given sets of training data to test each attribute at every node. It measures how well a given attribute separates the training examples according to their target classification. It uses statistical property call information gain to select which attribute to test at each node in the tree. Once a tree is built, it is applied to each tuple in the database and results in classification for that tuple. The basic idea is that all examples are mapped to different categories according to different values of the condition attribute set; its core is to determine the best classification attribute from the sets. Usually the attribute that has the highest information gain is selected as the splitting attribute of the current node. The ID3 algorithm had some challenges which were addressed by Ross Quinlan with the introduction of the C4.5 algorithm.

#### 2.2.1.2 C4.5

The C4.5 algorithm is Quinlan's extension of his own ID3 algorithm for generating decision trees. The C4.5 algorithm recursively visits each decision node, selecting the optimal split, until no further splits are possible and it is not restricted to binary splits, Larose (2005). For categorical attributes, C4.5 by default produces a separate branch for each value of the categorical attribute. Being a supervised learning algorithm, it requires a set of training examples and each example can be seen as a pair: input object and a desired output value (class). The algorithm analyzes the training set and builds a classifier that must be able to correctly classify both training and test examples, Larose (2005). A test example is an input object and the algorithm must predict an output value. The classifier used by C4.5 is a decision tree and this tree is built from root to leaves. It uses the concept of information gain or entropy reduction to select the optimal split. It is a well-known algorithm used to generate a decision trees and it is also an extension of the ID3 algorithm used to overcome its disadvantages. The decision trees generated by the C4.5 algorithm can be used for classification, and for this reason, C4.5 is also referred to as a statistical classifier. The C4.5 algorithm made a number of changes to improve ID3 algorithm. Some of these are:

- Handling training data with missing values of attributes
- Handling differing cost attributes
- Pruning the decision tree after its creation
- Handling attributes with discrete and continuous values

#### 2.2.1.3 Naive Bayes

The Naive Bayes algorithm (NB) can be used for both binary and multiclass classification problems. Naive Bayes algorithm builds and scores models extremely rapidly; it scales linearly in the number of predictors and rows. Naive Bayes algorithm makes predictions using Bayes' Theorem which derives the probability of a prediction from the underlying evidence. Bayes' Theorem states that the probability of event A occurring given that event B has occurred (P(A|B)) is proportional to the probability of event B occurring given that event A has occurred multiplied by the probability of event A occurring ((P(B|A)P(A)).Naïve Bayesian classifiers assume that there are no dependencies amongst attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, hence is called "naive". This classifier is also called idiot Bayes, simple Bayes, or independent Bayes. Some advantages of Naive Bayes are:

- It uses a very intuitive technique. Bayes classifiers, unlike neural networks, do not have several free parameters that must be set. This greatly simplifies the design process.
- Since the classifier returns probabilities, it is simpler to apply these results to a wide variety of tasks than if an arbitrary scale was used.
- It does not require large amounts of data before learning can begin.
- Naive Bayes classifiers are computationally fast when making decisions.

### 2.3 Effective use of school data

Many educational researchers have described education as a field in which practitioners make decisions based on intuition and gut instinct (Slavin, 2002). There are many research materials available describing the variety of ways in which data has supported educational decisions (Feldman & Tung, 2001; Lachat, 2002; Pardini, 2000; Protheroe, 2001). According to (Chrispeels, 1992; Earl & Katz, 2002) research on school improvement and effectiveness has shown data use to be central to the school improvement process. Data use is not a choice anymore for school authorities, but a must. This is because; data can be used to inform decisions on a wide variety of educational challenges. According to

Streifer (2002), “one of the many ways data can be used is by identifying the root cause of such challenges”. Chrispeels et al. (2000) showed data use to be a strong predictor of the efficiency of school improvement systems. Data use in schools does not only increase efficiency directly, but also serve as an intermediary for the positive effect of the other factors. For the purposes of this study, it will have been very significant if the schools had accumulated data in place to aid researchers in the academic evaluation process but public basic schools don’t lay much emphasis on the storage of such data thus making it difficult and less attractive to undertake data mining research in Ghanaian schools. Although there is an acute insufficiency of research conducted in this area, using school data to improve decision making is a good strategic tool for school administrators. School database systems draw upon many different types of information such as student performance data which is an integral part of the data-driven decision-making process. Supporters of data-driven decision-making practices argue that the effective use of school data enables schools systems to learn more about their schools, pinpoint successes and challenges, identify areas of improvement, and help evaluate the effectiveness of programs and practices (Mason, 2002). Kennedy (2003) included use of data as a central component of his model for raising achievement test scores. Earl and Katz (2002) noted that school authorities involved in the use of data often develop a mindset of being in charge of their own destiny and are increasingly able to find and use information to inform their school’s improvement. Feldman and Tung (2001) observed that schools involved in data use often evolved toward a more professional culture. Armstrong and Anthes (2001) also found that data use was helpful in raising teacher expectations of at-risk students, noting positive changes in teacher attitudes regarding the potential success of previously low performing students. However, the academic success of a school is measured by their academic performance. This is why schools are making good use of data to attain high level success rate. Student performance data can be used for various purposes, including evaluating the progress of students in their final external examinations, monitoring and improving students’ performance, determining where assessments converge and diverge, and judging the efficiency of the school curriculum and learning resources (Crommey, 2000). When school authorities become well informed on how to use data, they can effectively review their existing capabilities, identify weaknesses, and better chart plans for improvement (Earl & Katz, 2006). Research shows that data driven decision making has the potential to increase student performance (Alwin, 2002; Doyle, 2003; Johnson, 1999,

2000; Lafee, 2002; McIntire, 2002). Data need to be actively used to improve teaching and learning in schools but most schools often lack the capacity to implement what research suggests (Diamond & Spillane, 2004; Ingram et al., 2004; Mason, 2002; Petrides & Nodine, 2005;Wohlstetter, Van Kirk, Robertson, & Mohrman, 1997). The Ghana education service professes to play a significant role in helping schools build the necessary skills and capacity to use school data for effective decision making but according to Akyeampong (2008), 'the relatively low rates of return to public basic schools is also an indication that overall, the Ghana education service has been inefficient in preparing the large number of students who qualify for Senior high school or actively participate in the labour market'. Although the use of school data has proven to be beneficial to schools, the process of gathering and using such data is a difficult one. Computer systems have been used to support businesses and organizations for several years but it is a major challenge when it comes to the education sector. Thorn (2001) states that schools face technical challenges because of the variety of data they need to generate and use in their schools. This is because schools often have their data scattered across different locations thus making it difficult for its efficient organization. The advancement in technology is helping some schools to overcome these technological challenges. Stringfield et al. (2003) forecast that schools will soon have a variety of affordable and efficient computer tools to help in the data management process. It is therefore worthy to note that, high-performing schools make decisions based on data and not on instinct (Supovitz& Taylor, 2003; Togneri, 2003).

**2.4 Socio-economic factors and academic performance**

This section of the literature review provides some understanding of some of the socio-economic factors that influence academic performance. The choice of theories and factors reviewed here are based on their importance to the current study. At independence, many countries look to reform education to accelerate economic and social development. Ghana was no exception, and the newly independent government saw in education the keys to social and economic development. (Akyeampong, 2008). Ghana is said to be the first independent sub-Saharan African country outside South Africa to embark on a comprehensive drive to promote science education and the application of science in industrial and social development (Anamuah-Mensah, 1999). Public-funded institutions in Ghana are under scrutiny because of the recent global economic downturn which demands that developing countries improve efficiency in financial resource utilization. Government funded schools

can therefore not afford to remain unconcerned about the poor performance of public basic schools in the Basic education certificate examination. Research has shown that, social and economic factors have been one of the most studied causes in the academic evaluation of students. The poor performance of students in examinations in recent times could be attributed to the changing life pattern in some families coupled with the present economic hardship in the country which has made most families unable to meet their responsibilities of ensuring a healthy and literate family. A presidential commission on Education reforms in Ghana examined the reasons why most basic school students were unable to access senior secondary, and blamed this on a number of factors: inadequate facilities and infrastructure, parents unable to afford secondary fees, a lack of alternative tracks for students with different interests and abilities, an inability of students to meet the minimum requirements for further education and a lack of interest in further education (GOG, 2002). Akyeampong (2005) suggested that, teacher shortages in the technical/vocational subject areas effectively reduced quality of provision and undermined student interest. Studies investigating the impact of family size on academic performance show that family size such as the number of children has resource dilution hypothesis where the material resources and parental attention are diluted with additional children in the household, (Bachman, 2000). In a cross-country study testing the impact of family size on academic performance also concluded that, much of the association between family size and educational outcomes is simply due to the correspondence between large families and lower socioeconomic status (Marks, 2006). The size of the family in which a child grows affects his intellectual development; this is because in a large-size family, a child may not be given the required attention especially in his/her academics as the family will have more persons to cater for. The issue of payment of school fees, homework, attending Parent Teachers Associations (PTA) meetings and many more may not be convenient for the parents as they have to cater for many children. While children are well catered for and perform better in small-size family, a large family impacts negatively on the gross performance of the child. Family financial resources, which are associated with parents' occupation and educational attainment, often imply increased learning opportunities both at home and in school. Better-educated parents can contribute to their children's learning through their day-to-day interactions with their children and involving themselves in their children's school work. According to Osunloye, (2008) and Ushie et al., (2012), family background in terms of family type, size, socio-economic status and educational

background play important role in children's educational attainment and social integration. In a study, Heyman (1980) emphasized the importance of family income on pupils' performance. He opined children born into wealthier homes do better in many aspects of life and have high moral reasoning and better academic performance compared to children who come from poor homes. Maani (1990) observes that pupils' success at schools is closely related to their home backgrounds. These include; level of education of parents, family income, parents' marital status, and attitudes of parents towards education of their children and the children's attitudes and the quality of learners admitted in school.

It is also worthy to note that, parents are in the strongest position to develop positive relationships with their children that will facilitate the acquisition of standards and values; and also parents are better able to monitor and understand their children behavior than anyone else because of their long and sustained exposure to them, as a result to minimize the effect of a lower level of academic achievement and psychological problems (Deci et al, 1981). Nowadays, the condition of parents' marital status plays a great role for child development. Specifically, divorce and remarriage have an impact on children adjustment. For this reason, there is an increased attention being given by scientists to investigate the overall effect of family transitions on the well being of children. Regarding the relationship between family conditions and the behavior of children, some studies have shown that harmonious marriages promote children's competence and maturity. Others demonstrate that marital conflict tends to be associated with the children's cognitive delay, school difficulties, and antisocial or withdrawn behavior in the early school years (Bond & Mcmahon, 1984; Emery, 1982; Gottman & Katz, 1989; Hetherington, Cox and Cox, 1979; 1986; Rutter, 1978; Wallerstein & Kelly, 1980; Weissman, 1983; Brown 2004; Fomby and Cherlin 2007). Raychauduri et al. (2010), also states that, socio-economic factors like class attendance, family income, and mother's and father's education, teacher-student ratio, presence of trained teacher in school, sex of student and distance of school affect the performance of students. Hijaz and Naqvi (2006) also conducted a study on student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab university of Pakistan. The hypothesis stated was "Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' family income, students' mother's age and mother's education are significantly related with student performance". By means of simple linear regression analysis, it was found that, factors like mother's education and student's family income were highly

correlated with the student academic performance. According to Hansen and Mastekaasa (2006), the cultural capital theory expects students from families who are closest to the academic culture to have greatest success. It is also believed that low socio-economic status of students' affects their academic progression negatively because it prevents access to vital learning resources and creates additional stress at home. (Eamon 2005; Jeynes, 2002). A number of studies have been carried out to identify and analyse the numerous factors that affect academic performance in various academic institutions. Most of their findings identify parents' education, family income (Devadoss& Foltz, 1996), self-motivation, age of student, learning preferences Aripin, et al(2008), class attendance (Romer, 1993), and entry qualifications as factors that have a significant effect on the students' academic performance in various situations. Other studies have also noted that, the academic achievement of students is based upon the parent's socio-economic standing in society. According to a research conducted by Acheampong et al.,(2004) 'relationship to the household head is found to affect educational attainment significantly. Children of the household head are most likely to have made educational progress than servants'. They also concluded that, greater proportions of children under the age of 7 in a household are found to reduce the probability of a household member reaching post-primary education. Much more important was the occupational or socio-economic status of the household head and they concluded that, household members in a household with a head in formal public or private sector employment are much more likely to have progressed beyond junior high school than in households headed by a food farmer all things being equal. Public basic school students in Ghana go through a lot of socio-economic difficulties that go a long way to affect their academic performance. Students walk long distances every morning to attend school, they get to school tired and have difficulty concentrating in class. Some are made to sell in the morning before going to school thus making it difficult for them to get to school early and prepare for class. Others are made to sell late into the night before retiring to bed; denying them of the much needed time to go through the day's lessons. Sogbetun (1981) and Hassan (1983) have examined the causes of poor academic performance among secondary school students. Some of the factors identified are intellectual ability, poor study habit, achievement motivation, lack of vocational goals, low self-concept, low socio-economic status of the family, poor family structure and anxiety. Poor academic performance takes many forms and according to Acheampong (2008), 'one of the reasons was that the quality of practical education students received depended on whether they

attended a school in a rural or urban area'. Mulkey et al. (1992) reported that, while family income is important, other factors have a greater influence on academic performance. They suggested that parental expectations, family size, and the quality of the parent-child relationship are stronger predictors of future academic success than income. According to different scholars, children from intact families, on average, display less behavior problems, less psychological distress, and greater academic achievement than do children of divorce (Allison and Furstenberg, 1989; Guidubaldi and Perry, 1985; Hetherington et al., 1985; Wallerstein et al., 1988;Gershoff, Aber, Raver, and Lennon 2007). Somewhat differently, other studies have also sought to examine the importance of the access to both parents than separated. In one study following parental separation, in general, 30% of the children has experienced a marked decrease in their academic performance, and this was evident three years later (Bisnaire et al., 1990;, Aughinbaugh, Pierret, and Rothstein 2005). Access to both parents seemed to be the most protective factor, in that it was associated with better academic adjustment. Moreover, the data revealed that non-custodial parents (mostly fathers) were very influential on their children's development. These data also support the interpretation that the more time a child spends with the non-custodial parent, the better the overall adjustment of the child. (Graetz, 1995) did a study on socio-economic status of the parents of students and came to the conclusion that, the socio-economic background of parents has a great impact on student's academic achievement, serves as the main source of educational imbalance among students. Considine and Zappala (2002) in their study on the influence of social and economic challenges in the academic performance of students also noticed that, children who come from low income families show low retention rates, low literacy level, have behavioral problems in school, difficulties in their studies and mostly display negative attitude towards studies and school and this affects their test scores in examination. McMillan and Westor (2002) argued that social economic status is comprised of three major dimensions: education, occupation and income and therefore in developing indicators appropriate for high education context, researchers should study each dimension of social economic status separately. They added that education, occupation and income are moderately correlated therefore it is inappropriate to treat them interchangeably in the higher education context. According to Jeynes(2002), "Social and economic status of students is generally determined by combining parents' qualification, occupation and income standard". But it is interesting to note that, Pedrosa et.al. (2006) had a different opinion. In their study on social and

educational challenges among school children, they noted that students who mostly come from deprived socio-economic backgrounds perform relatively better than those coming from higher socio-economic backgrounds. Determining the most contributing factors in the evaluation of academic performance that can be applied to all situations is a very complex and challenging job. This is because students everywhere belong to a variety of backgrounds depending on where they live. This diversity is vast and complex in a multi cultured country like Ghana. It is therefore important to note that, the criterion for categorizing socio-economic standards in different countries is relative and the impact of these factors varies in terms of the extent and direction. The criteria for categorizing low socio-economic status for a developed country like Germany will be different from the criteria used in a developing country like Ghana. Escarce (2003) pointed out that residential stratification and segregation has ensured that students belonging to low-income backgrounds usually attend schools with lower funding levels, and this situation has led to a reduction in academic achievement and poor motivation of the students. Such students also stand a high risk of educational malfunction in future endeavors. An additional challenge is the rising cost of secondary education to both government and parents and the potential that this has on constraining future growth (Akyeampong 2005). Sentamu (2003) argued that, schools influence educational process in content organization, teaching and learning. He argued that rural families and urban families where both parents were illiterate do not seem to consider home study for their children a priority and that such illiterate families will not foster a study culture in their children since the parents themselves did not attend school or the education they received was inadequate to create this awareness in them. These differences in home literacy activities reflect in the academic performance of the children. Competition, selection and choice has also began to take root in primary education which has limited access to secondary education, especially for children from disadvantaged and poor households (Addae-Mensah, Djangmah & Agbenyega, 1973). But Recent events shows that, private sector provision of basic and secondary education has grown, some would say, offering more choice for families than ever before. However, there is growing evidence that it might also be acting as a tool for social mobility and stratification in Ghanaian society (Addae-Mensah 2000; Donge et al., 2003). According to Acheampong et al., (2004). ‘the end of basic education marks a key transition in education in Ghana, since it is post-basic education which is associated with substantial and increasing economic returns to schooling and with selection into more

lucrative occupations, especially wage-employment in both the public and private sectors'. In Ghana, students from the private basic schools are expected to excel at the Basic Education Certificate Examination because they have access to quality teaching and learning resources, well supervised teachers, parents willingness to pay extra money to motivate the teachers and the recruitment of part-time teachers by parents to assist their children at home. Crosne et al. (2004) noticed that school ownership, provision of facilities and availability of resources in school is an important structural component of the school.

**2.5 Review of literature on the test variables**



The influence of socio-economic factors on academic performance has been investigated in a number of studies with widely differing conclusions. Most of the differences in reported findings are due to varying contexts in which the study is conducted. In the light of the theories, techniques, methods and models reviewed in the preceding literature, especially the insights from Baker and Yacef (2009); Chrispeels (1992); Earl & Katz (2002); Jeynes (2002); Eamon (2005); Greatz (1995); Considine&Zappala (2002); Hansen &Mastekaasa (2003).Conclusions can be drawn that, the cultural and economic conditions under which they conducted their studies were different. It is therefore worthy to note that the results of these studies could not be subjected to generalizations beyond this cultural and economic environment. This is also largely because, the criteria for categorizing low socio-economic families in a developed economy is relatively different from the criteria used in a developing economy. (Bachman, 2000) concluded in his studies that, large family size puts resource constraints on parents. The research conducted by (Marks, 2006) also confirms Bachman's findings and re-affirms the fact that, the impact of large family size on academic performance is negative. Research findings from Osunloye, (2008) and Ushie et al., (2012), espoused that, family type, family size, educational background, social and economic status play a very important role in children's educational success. Heyman (1980) also emphasized the importance of family income on students' performance. (Maani, 1990) researched on the level of education of parents, family income, parents' marital status, parent's attitude towards the education of their children. Raychauduri et al. (2010), researched on socio-economic factors like class attendance, family income, and mother's and father's education and distance of school. Hijaz and Naqvi (2006) also conducted a study on students' family income and mother's education. Sogbetun (1981) and Hassan (1983) examined the causes of poor

academic performance by looking at low socio-economic status of the family and poor family structure. Mulkey et al. (1992) also researched on family size, and the quality of the parent-child relationship. McMillan and Westor (2002) argued that social economic status comprised of three major dimensions: education, occupation and income. Looking at the variables that have been researched on by all these researchers, it inevitable means that, the socio-economic factors of poor academic performance cannot be investigated thoroughly without considering the size of the family, parents educational background, parents employment status, parent's marital status and family income. These factors were chosen for the current study because similar works has been conducted in other countries and the findings from these works backed the factors chosen for the current study. It is also worthy to note that, left for most researchers to choose which socio-economic factors may affect academic performance, these factors will be chosen. This is based on the views espoused by many researchers who conducted similar research works. Investigating these factors therefore helps to address the research objectives. This study proposes to apply the C4.5 algorithm to analyze these socio-economic factors in the Ghanaian context. The C4.5 algorithm is chosen for this study because it is a well-known decision tree induction learning technique that has been successfully and extensively applied to educational data. It is a better algorithm compared to ID3 because it address the problem of over fitting data; error pruning, rule post-pruning, continuous attributes and handling of missing attribute values. It is also a well-known algorithm used for classifying datasets by inducing decision trees and rules from datasets which could contain categorical and numerical attributes. This perfectly fits this study because the training data set used for this study contains both categorical and numerical attributes. The rules derived could be used to predict categorical values of attributes from new records. The attractiveness of this tree-based method is due to the fact that, in contrast to neural networks, decision trees represent rules which can readily be expressed in a language that humans can understand. Kotsiantis et al. (2003) used only demographic data to classify university students into dropouts and non-dropouts. Using only demographic data can be a bit problematic since it might not reflect the true picture on the ground. Minaei-Bidgoli et al. (2003) also conducted a study where they tried to predict the final test grades for students enrolled in a web-based course. Several learning algorithms were compared: decision trees, neural networks, naïve Bayes, logistic regression, support vector machines, and KNN. According to Zou, and Huang, (2005) KNN is well suited in situations where predictions need to be made from noisy

and incomplete data but because the noise level of my training data set is very low, the C4.5 will be the most ideal algorithm to use. El-Halees (2008), Radaideh (2006) and Baradwaj (2011) used educational data mining to analyze students' learning behavior, study the main attributes that may affect student performance in courses and apply classification as a data mining technique to evaluate student' performance. They also used the decision tree method to conduct the classification but their studies focused on classroom variables. This study seeks to look beyond the classroom and focus on socio-economic challenges faced by such students. It seeks to determine whether there is a correlation between the selected socio-economic factors and academic performance. One disadvantage to Romero and Ventura's (2010) studies on EDM is that the results are not necessarily generalizable to other institutions. This means that the results are highly associated with a specific educational institution. Chandra and Nandhini (2010) applied association rule based on students' failed courses to identify their failure patterns. One weakness to this algorithm is that, it has too many parameters for somebody non expert in data mining and the obtained rules are far too many, most of them non-interesting and with low comprehensibility. An experiment by Aman and Suruchi, (2007) conducted in a WEKA (Waikato Environment for Knowledge Analysis) environment by using three algorithms namely ID3, C4.5, and Simple CART revealed that; the C4.5 classifier outperforms the rest in terms of classification accuracy. This is positive for this study because inaccurate classification can lead to wrong interpretations.



**2.6 Conceptual framework for the academic performance classification model**

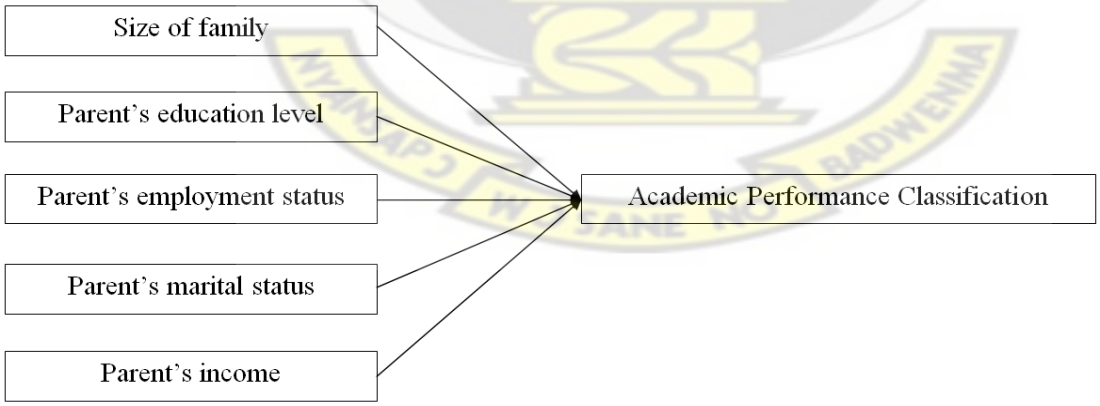


Figure 1.1 shows the underlining theory and concepts of the research.

**2.7 Research hypotheses in data mining**

Hand, Mannila and Smyth (2001) described exploratory data analysis as data-driven hypothesis generation. They indicated that, data is examined in search of structures that may indicate deeper relationships between cases or variables. This process stands in contrast to hypothesis testing which

begins with a proposed model or hypothesis and undertakes statistical manipulations to determine the likelihood that the data arose from such a model. According to Hand, Mannila and Smyth (2001) “In data mining, it is the patterns in the data that give rise to the hypotheses in contrast to situations in which hypotheses are generated from theoretical arguments about underlying mechanisms”. Data mining is primarily concerned with looking for unsuspected features as opposed to testing specific objectives that are formed before we see the data. They believe that in practice, data mining algorithm generates potentially interesting hypothesis that would need to be explored further. A standard machine learning technique is to separate the set of examples into a training set and a test set. Myers and Walpole (1978) stated that, ‘the standard way to evaluate a hypothesis in supervised learning is to split the data into two portions; the training set and the test set. The training set is used in order to produce hypotheses, and the test set, which is never seen during the hypothesis forming stage is used to test the accuracy of the hypothesis in predicting the categorization of unseen examples’. In this way, we can have more confidence that the learned hypothesis will be of use to us when we have a genuinely new example for which we do not actually know the categorization. The interdependency of one or more socio-economic variables cannot be ruled out in this study. This is because rules generated from the decision tree make it imperative for one or more variables to be tested together. The testing is focused on the accuracy of the model after it has been developed by running the model with a test set of data to measure how accurately the model can perform on an unseen data.

# CHAPTER 3

# RESEARCH METHODOLOGY

## 3.0 Introduction

This chapter provides information on the research methods of this thesis. The survey research method has been chosen to analyze the factors influencing academic performance in public basic schools. The study population and sample size has been described followed by the sampling techniques and research instrument. Reliability and validity, measurement procedures, and the CRISP-DM (cross industry process for data mining) are discussed. The survey instrument has been designed using Likert categorical scale to measure respondents' attitude towards the socio-economic factors consistent to previous research findings. The data collection and ethical considerations are also included in this chapter.

## 3.1 Study population and sample size

According to Burns and Grove (1993), 'a population is defined as all elements (individuals, objects and events) that meet the sample criteria for inclusion in a study'. The study population consisted of 800 public basic students in the Ablekuma west constituency of the Greater Accra region. Polit & Hungler (1993) describes a convenient sample as subjects included in a study because they happen to be in the right place at the right time. A convenient sample of 200 subjects was selected from 10 public basic schools. This represents 25% of the population hence 200 questionnaires were distributed to the participants. Mouton (1996) defines a sample as elements selected with the intention of finding out something about the total population from which they are taken.

## 3.2 Sampling technique

The Systematic random sampling technique is employed in selecting the sample from the targeted population. A systematic random sampling is a type of probability sampling method in which sample members from a large population are selected according to a random starting point and a fixed periodic interval. In order for systematic sampling to work effectively, the units in the population is randomly ordered, at least with respect to the characteristics being measured. The process starts with an assumption that, N units of the population are numbered 1 to N in some order. To select a sample of n units, we must take a unit at random from the first k units and every kth unit thereafter. Cochran (1953) & Yamane (1967 ). This sampling technique is chosen because of its simplicity and its flexibility to allow the researcher to add a degree of processes into the random selection of subjects. It is fairly easy to do and is widely used for its convenience and time efficiency, Cochran (1953). In many surveys, it is

found to provide more precise estimates than simple random sampling, Raj (1972) & Cochran (1953). This happens when there is a trend present in the list with respect to the characteristic of interest. The population for this study is made up of N=800 students and they are listed in a random order. A sample size of n=200 students is required for this study. The sampling fraction is n/N = 200/800 = 25. In this case, the interval size is equal to N/n = 800/200 = 4. A random integer is selected starting with the 4th unit in the list and take every 4th unit (because k=4).

### 3.3 Research instrument

The study uses questionnaire as the main data-gathering instrument. (See Appendix A). Close ended questionnaires were used because the population is large and time for collecting data is limited. The research design is of crucial importance because it determines the success or failure of research. The research design guides logical arrangements for the collection and analysis of data so that conclusions may be drawn. The questionnaire was divided into two main sections: a profile and the survey proper. The profile contains socio-demographic characteristics of the respondents such as age, gender and class. The survey proper explored the perceptions of students on the socio-economic challenges they face. The questions were structure using the Likert format. In this survey type, four choices are provided for every question or statement. The choices represent the degree of agreement each respondent has on the given question. The scale below shows the range Interpretation used to interpret the total responses of all respondents for every survey question by computing the weighted mean.

Table 3.1 Likert scale range interpretation

| Scale | Interpretation |
|---|---|
| 4 | Strongly agree |
| 3 | Agree |
| 2 | Disagree |
| 1 | Strongly disagree |

The Likert survey was the selected questionnaire type as this enabled the respondents to answer the survey easily. In addition, this research instrument allowed the research to carry out the quantitative approach effectively with the use of statistics for data interpretation.

### 3.3.1 Research Design

In order to examine the effect of socio-economic challenges on academic performance, descriptive research design was used. A descriptive survey is selected because it provides an accurate portrayal of the variables involved and the ability to collect data from a large group within a short period. This design has been chosen to meet the objectives of the study. According to Mouton (1996), a descriptive survey is used to collect original data for describing a population too large to observe directly. This questionnaire, which is used as a quantitative data-collection instrument, has the objective of collecting certain demographic and socio-economic information. The questions are organized into two sections. Section A of the survey includes demographic variables such as age, gender and class and section B includes six multi dimensional constructs that reflect the main issues affecting students' academic performance in public basic schools. The self-designed questionnaire comprise of the following:

- Age of student
- Gender of student
- Class of student
- Size of family
- Education level of parent
- Employee status of parent
- Marital status of parent
- Income level of parent
- Academic performance of student

### 3.4 Data collection

The study collected both primary and secondary data. Primary data was collected through questionnaires which contained closed ended questions. The closed ended questions were expected to facilitate the collection of accurate information. Data was collected from 20 students from each of the 10 selected public basic schools. A total of 200 questionnaires were sent out and 177 questionnaires were received out of which 156 were found to be fully filled in. The rest of 21 questionnaires were discarded due to incomplete information. Secondary data was collected relying on reviews from national and international journals as well as research literature. The literature and related journals was collected from various sources such as internet, library references and any other relevant online database.

### 3.5 Demographic relationships and study variables

Although it was not part of the purpose of the study, this set of data was intended to describe demographic variables of the sample. Participants were asked to indicate their age, gender and class. The tables below show the various demographic statistics and their percentages.

Table 3.2 Association between age and respondents

| Age in years | Total Respondents | Percent |
|---|---|---|
| 13 | 59 | 37.82% |
| 14 | 68 | 43.59% |
| 15 | 29 | 18.59% |



Table 3.3 Association between Gender and respondents

| Gender | Total Respondents | Percent |
|---|---|---|
| Male | 72 | 46.2% |
| Female | 84 | 53.8% |

Table 3.4 Association between class and respondents

| Class | Total Respondents | Percent |
|---|---|---|
| JHS 1 | 81 | 51.9% |
| JHS 2 | 75 | 48.1% |

### 3.6 Reliability and validity

Polit and Hungler (1993) refer to reliability as the degree of consistency with which an instrument measures the attribute it is designed to measure. The questionnaire which was answered by the students revealed consistency in responses. Reliability was also ensured by minimizing sources of measurement error like data collector bias. This error type was minimized because the researcher was the only one to administer the questionnaires, and thereby standardizing conditions such as exhibiting friendliness and support. According to Polit and Hungler (1993), 'The validity of an instrument is the degree to which an instrument measures what it is intended to measure'. In order to test the validity of the research tool which used for this study, the researcher tested the questionnaire to 5 respondents. These respondents as well as their answers were not part of the actual study process and were only used for testing

purposes. After the questions have been answered, the researcher asked the respondents for any suggestions or any necessary corrections to improve the instrument further. The researcher modified the content of the questionnaire based on the assessment and suggestions of the sample respondents.

### 3.7 Measurement Procedures

The questions on the questionnaire are formulated based on the objectives, questions and hypothesis of this research. The questions will follow a logical progression to sustain the interest of respondents and gradually stimulate question answering. The Likert categorical scale will be used to measure the respondents' multi-dimension constructs measurement. Likert scale measures are commonly used in assessing student performance as well as student

perceptions. They are particularly good in gathering data to subjective questions. The measurement procedure is undertaken by analyzing the responses by assigning weighting to the responses. For a positive statement the response indicating the most favorable measurement is given the highest score. For the four-category scale. 4 is assigned to the most favorable measurement "strongly agree" and 1 is assigned to the least favorable measurement "strongly disagree".

#### 3.7.1 Why Likert scale

Over the years, numerous methods have been used to measure character and personality traits Likert (1932). A variety of methods are available to assist evaluators in gathering data. One of those methods involves the use of a scale. The Likert-type scale involves a series of statements that respondents may choose from in order to rate their responses to evaluative questions (Vogt, 1999). The difficulty of measuring attitudes, character, and personality traits lies in the procedure for transferring these qualities into a quantitative measure for data analysis purposes. Likert (1932) responded by developing a procedure for measuring attitudinal scales. The original Likert scale used a series of questions with five response alternatives. He combined the responses from the series of questions to create an attitudinal measurement scale. His data analysis was based on the composite score from the series of questions that represented the attitudinal scale. While Likert used a five-point scale, other variations of his response alternatives are appropriate, including the deletion of the neutral response (Clason & Dormody, 1994). Likert data are generated as ordinal direct response data. With Likert data, agreement with a statements is measured at an ordinal level. The Likert measurement model is friendly to deploy and remains one of the popular models because of the following;

- The method is based entirely on empirical data regarding subjects' responses rather than the subjective opinions of judges.

- This method produces more homogeneous scales and increases the probability of a unitary attitude being measured; as a result, validity (construct and concurrent) and reliability are reasonably high; and greater ease of preparation.

### 3.7.2 Analyzing likert response items

The Likert scale was used to interpret items in the questionnaire. These responses were based on the respondents' rating of the socio-economic factors affecting academic performance. The range and interpretation of the four-point scale is shown in Table 3.5

Table 3.5 The four point likert scale

| Scale | Range | Interpretation |
|---|---|---|
| 4 | 3.01 – 4.00 | Strongly agree |
| 3 | 2.01 – 3.00 | Agree |
| 2 | 1.01 - 2.00 | Disagree |
| 1 | 0.00 - 1.00 | Strongly disagree |

After gathering all the completed questionnaires from the respondents, total responses for each item were obtained and tabulated. In order to use the Likert-scale for interpretation, weighted mean to represent each question was computed. Weighted mean is the average wherein every quantity to be averages has a corresponding weight. These weights represent the significance of each quantity to the average. To compute for the weighted mean, each value must be multiplied by its weight. Products would then be added to obtain the total value. The total weight would also be computed by adding all the weights. The total value is then divided by the total weight. Statistically, the weighted mean is calculated using the following formula;

$$\bar{x} = \frac{\sum fi\ xi}{\sum fi}$$

Where $\bar{x}$ = weighted mean

$x_i = x_1, x_2, x_3$……= number of responses

$f_i = f_1, f_2, f_3$ ……= frequencies corresponding to the given items

The tabulation of student response data is presented in the table below

*Table 3.6 Tabulation of student response data*

| Q | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 4 | 3 | 4 | 2 | 3 | 2 | 4 | 4 | 3 | 2 | 3 | 2 | 4 | 2 | 4 | 4 | 2 | 3 | 1 | 3 | 3 | 2 | 2 | 4 | 2 | 3 | 2 | 3 | 3 | 2 | 4 | 1 | 4 | 2 | 4 | 2 | 2 | 4 | 3 | 3 | 3 | 2 | 4 | 3 | 2 | 1 | 3 | 3 | 2 |
| 2 | 2 | 1 | 4 | 2 | 3 | 2 | 3 | 4 | 4 | 2 | 3 | 4 | 1 | 3 | 2 | 3 | 2 | 3 | 4 | 2 | 4 | 1 | 3 | 2 | 4 | 4 | 2 | 3 | 2 | 3 | 2 | 4 | 4 | 2 | 3 | 3 | 2 | 2 | 4 | 1 | 2 | 3 | 4 | 2 | 3 | 2 | 2 | 4 | 2 | 3 |
| 3 | 3 | 2 | 4 | 2 | 3 | 3 | 4 | 2 | 4 | 4 | 2 | 3 | 2 | 4 | 2 | 3 | 1 | 3 | 3 | 1 | 4 | 2 | 4 | 3 | 1 | 4 | 2 | 4 | 3 | 2 | 3 | 1 | 3 | 4 | 2 | 2 | 4 | 2 | 3 | 3 | 2 | 2 | 4 | 4 | 2 | 3 | 2 | 1 | 4 | 2 |
| 4 | 4 | 1 | 3 | 2 | 3 | 4 | 4 | 2 | 3 | 2 | 3 | 2 | 3 | 1 | 4 | 4 | 2 | 3 | 2 | 3 | 4 | 2 | 2 | 4 | 3 | 4 | 3 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 3 | 1 | 3 | 1 | 4 | 2 | 4 | 1 | 3 | 1 | 3 | 2 | 4 | 2 | 4 | 2 |
| 5 | 4 | 2 | 4 | 2 | 3 | 2 | 2 | 3 | 1 | 4 | 2 | 3 | 3 | 1 | 1 | 2 | 4 | 2 | 3 | 3 | 2 | 2 | 3 | 2 | 4 | 3 | 1 | 4 | 2 | 2 | 4 | 2 | 4 | 2 | 4 | 2 | 3 | 2 | 3 | 3 | 2 | 2 | 4 | 1 | 2 | 4 | 4 | 2 | 2 | 4 |
| 6 | 3 | 2 | 2 | 4 | 2 | 3 | 3 | 2 | 3 | 2 | 4 | 2 | 2 | 2 | 4 | 4 | 2 | 2 | 2 | 3 | 3 | 2 | 3 | 2 | 3 | 4 | 3 | 2 | 2 | 2 | 3 | 2 | 2 | 3 | 2 | 3 | 2 | 4 | 2 | 3 | 3 | 2 | 4 | 2 | 4 | 2 | 4 | 2 | 3 | 2 |

| 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 2 | 4 | 4 | 2 | 3 | 2 | 3 | 3 | 1 | 4 | 4 | 2 | 4 | 3 | 1 | 2 | 4 | 2 | 3 | 3 | 4 | 4 | 2 | 3 | 2 | 2 | 4 | 3 | 3 | 1 | 3 | 2 | 3 | 2 | 2 | 3 | 2 | 2 | 4 | 1 | 2 | 3 | 2 | 3 | 2 | 4 | 2 | 3 | 2 |
| 3 | 1 | 4 | 4 | 2 | 2 | 3 | 2 | 3 | 2 | 4 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 4 | 4 | 2 | 4 | 1 | 3 | 2 | 2 | 3 | 1 | 4 | 4 | 2 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 4 | 1 | 2 | 3 | 2 | 3 | 4 | 1 | 3 | 4 | 2 | 2 |
| 3 | 2 | 4 | 4 | 2 | 4 | 3 | 2 | 3 | 3 | 4 | 2 | 4 | 2 | 3 | 1 | 4 | 2 | 3 | 2 | 3 | 1 | 4 | 2 | 4 | 4 | 2 | 2 | 3 | 3 | 1 | 4 | 2 | 4 | 2 | 3 | 2 | 4 | 2 | 4 | 2 | 3 | 1 | 3 | 2 | 4 | 2 | 3 | 2 | 3 |
| 4 | 2 | 3 | 2 | 3 | 3 | 3 | 4 | 3 | 1 | 2 | 4 | 1 | 4 | 3 | 2 | 3 | 2 | 4 | 4 | 1 | 3 | 1 | 3 | 1 | 4 | 4 | 2 | 3 | 1 | 3 | 4 | 4 | 4 | 4 | 2 | 3 | 2 | 3 | 2 | 4 | 2 | 1 | 4 | 2 | 4 | 2 | 3 | 2 | 3 |
| 2 | 4 | 2 | 4 | 2 | 4 | 2 | 2 | 3 | 4 | 2 | 4 | 1 | 3 | 2 | 2 | 3 | 2 | 4 | 2 | 4 | 2 | 1 | 3 | 2 | 2 | 4 | 2 | 4 | 2 | 3 | 2 | 2 | 2 | 3 | 2 | 4 | 2 | 2 | 1 | 4 | 2 | 2 | 2 | 3 | 1 | 2 | 4 | 2 | 3 |
| 3 | 3 | 2 | 3 | 2 | 3 | 2 | 3 | 3 | 2 | 2 | 2 | 4 | 2 | 4 | 2 | 4 | 2 | 3 | 2 | 2 | 3 | 2 | 3 | 2 | 2 | 3 | 2 | 2 | 4 | 2 | 3 | 2 | 2 | 3 | 2 | 3 | 2 | 2 | 4 | 2 | 4 | 4 | 2 | 2 | 3 | 2 | 3 | 2 | 2 |

| 101 | 102 | 103 | 104 | 105 | 106 | 107 | 108 | 109 | 110 | 111 | 112 | 113 | 114 | 115 | 116 | 117 | 118 | 119 | 120 | 121 | 122 | 123 | 124 | 125 | 126 | 127 | 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 | 141 | 142 | 143 | 144 | 145 | 146 | 147 | 148 | 149 | 150 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 3 | 2 | 1 | 3 | 2 | 4 | 1 | 4 | 3 | 1 | 1 | 3 | 2 | 3 | 1 | 1 | 3 | 2 | 4 | 3 | 2 | 4 | 2 | 2 | 4 | 1 | 4 | 3 | 2 | 1 | 3 | 1 | 4 | 2 | 4 | 3 | 1 | 3 | 2 | 1 | 4 | 2 | 4 | 2 | 3 | 1 | 3 | 1 | 2 |
| 3 | 3 | 2 | 4 | 1 | 3 | 4 | 1 | 3 | 2 | 3 | 2 | 4 | 1 | 4 | 4 | 2 | 3 | 2 | 4 | 3 | 3 | 2 | 4 | 4 | 1 | 3 | 4 | 1 | 3 | 1 | 4 | 4 | 1 | 3 | 2 | 3 | 2 | 3 | 2 | 4 | 4 | 1 | 3 | 1 | 4 | 1 | 4 | 3 | 3 |
| 4 | 1 | 3 | 3 | 2 | 4 | 2 | 3 | 3 | 1 | 3 | 1 | 4 | 2 | 4 | 3 | 3 | 2 | 4 | 4 | 3 | 2 | 3 | 3 | 2 | 4 | 3 | 1 | 3 | 3 | 1 | 4 | 2 | 4 | 1 | 4 | 1 | 3 | 4 | 1 | 3 | 3 | 4 | 2 | 4 | 3 | 3 | 1 | 4 | 4 |
| 4 | 2 | 3 | 3 | 2 | 4 | 4 | 2 | 3 | 1 | 2 | 3 | 3 | 4 | 2 | 3 | 3 | 2 | 4 | 1 | 3 | 2 | 3 | 2 | 3 | 2 | 4 | 1 | 3 | 2 | 3 | 1 | 3 | 2 | 4 | 3 | 2 | 4 | 1 | 4 | 1 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 4 | 1 |
| 2 | 2 | 4 | 1 | 3 | 3 | 2 | 4 | 2 | 4 | 3 | 1 | 2 | 4 | 2 | 2 | 4 | 1 | 3 | 2 | 2 | 3 | 1 | 4 | 2 | 4 | 1 | 4 | 3 | 2 | 1 | 2 | 4 | 2 | 4 | 1 | 4 | 2 | 3 | 3 | 4 | 1 | 3 | 2 | 4 | 1 | 4 | 3 | 2 | 4 |
| 3 | 2 | 4 | 2 | 4 | 3 | 3 | 1 | 4 | 2 | 2 | 3 | 2 | 4 | 2 | 2 | 4 | 1 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 4 | 1 | 3 | 3 | 3 | 4 | 4 | 2 | 3 | 2 | 2 | 2 | 2 | 4 | 3 | 3 | 4 | 2 | 4 | 2 | 3 | 3 | 2 | 4 | 2 |

*Table 3.7 The composite score from the series of questions that represented the attitudinal scale*

| Question | Total |
|---|---|
| 1 | **395** |
| 2 | **404** |
| 3 | **414** |
| 4 | **400** |
| 5 | **395** |
| 6 | **400** |

| Question | frequency (*fi*) | | | | No of responses (*xi*) | *fi xi* | | | ∑*fi* | ∑*fi xi* | $\bar{x}=\frac{\sum fi\,xi}{\sum fi}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | | | | | | | |
| 1 | 20 | 50 | 45 | 35 | 1 2 3 4 | 20 | 100 | 135 | 140 | 150 | 395 | 2.63 |
| 2 | 19 | 47 | 45 | 39 | 1 2 3 4 | 19 | 94 | 135 | 156 | 150 | 404 | 2.69 |
| 3 | 18 | 43 | 46 | 43 | 1 2 3 4 | 18 | 86 | 138 | 172 | 150 | 414 | 2.76 |
| 4 | 20 | 46 | 48 | 36 | 1 2 3 4 | 20 | 92 | 144 | 144 | 150 | 400 | 2.67 |
| 5 | 17 | 63 | 28 | 42 | 1 2 3 4 | 17 | 126 | 84 | 168 | 150 | 395 | 2.63 |
| 6 | 3 | 73 | 45 | 29 | 1 2 3 4 | 3 | 146 | 135 | 116 | 150 | 400 | 2.67 |

### 3.7.3 Data interpretation for training set

The following data interpretation represents the data used for the training set in the data mining analysis. For the purposes of this research, 100 responses were used. The following represents the percentage of responses by participants.

3.7.3.1 **The Socio-economic factors and academic performance attitude scale**

1. I come from a large family background?
2. My parents are educated?
3. My parents are gainfully employed?
4. My parents are married?
5. My parents are high income earners?
6. My academic performance can be rated as good?

The following results were obtained indicating a positive and negative response attitude:

- **Question 1**

  44% of the participants (n = 44) either agreed or strongly agreed that they come from small families.

  56% of the participants (n = 56) either disagreed or strongly disagreed that they come from small families.

- **Question 2**

  53% of the participants (n = 53) either agreed or strongly agreed that their parents are educated.

  47% of the participants (n = 47) either disagreed or strongly disagreed that their parents are educated.

- **Question 3**

  56% of the participants (n = 56) either agreed or strongly agreed that their parents are employed.

  44% of the participants (n = 44) either disagreed or strongly disagreed that their parents are employed.

- **Question 4**

  57% of the participants (n = 57) either agreed or strongly agreed that their parents are married.

  43% of the participants (n = 43) either disagreed or strongly disagreed that their parents are married.

- **Question 5**

  45% of the participants (n = 45) either agreed or strongly agreed that their parents are high monthly income earners.

  55% of the participants (n = 55) either disagreed or strongly disagreed that their parents are high monthly income earners.

- **Question 6**

  46% of the participants (n = 46) either agreed or strongly agreed that their academic performance is good.

  54% of the participants (n = 54) either disagreed or strongly disagreed that their academic performance is good.



### 3.7.4 Data interpretation for test set

The following data interpretation represents the data used for the test set in the data mining analysis. For the purposes of this research, 50 responses were used. The following represents the percentage of responses by participants.

3.7.4.1 **The socio-economic factors and academic performance attitude scale**

1. I come from a large family background?
2. My parents are educated?
3. My parents are gainfully employed?
4. My parents are married?
5. My parents are high income earners?
6. My academic performance can be rated as good?

The following results were obtained indicating a positive and negative response attitude:

- **Question 1**

  52% of the participants (n = 26) either agreed or strongly agreed that they come from small families.

  48% of the participants (n = 24) either disagreed or strongly disagreed that they come from small families.

- **Question 2**

  62% of the participants (n = 31) either agreed or strongly agreed that their parents are educated.

  38% of the participants (n = 19) either disagreed or strongly disagreed that their parents are educated.

- **Question 3**

  66% of the participants (n = 33) either agreed or strongly agreed that their parents are employed.

  34% of the participants (n = 17) either disagreed or strongly disagreed that their parents are employed.

- **Question 4**

  56% of the participants (n = 28) either agreed or strongly agreed that their parents are married.

  44% of the participants (n = 22) either disagreed or strongly disagreed that their parents are married.

- **Question 5**

  50% of the participants (n = 25) either agreed or strongly agreed that their parents are high monthly income earners.

  50% of the participants (n = 25) either disagreed or strongly disagreed that their parents are high monthly income earners.

- **Question 6**

  28% of the participants (n = 28) either agreed or strongly agreed that their academic performance is good.

  22% of the participants (n = 22) either disagreed or strongly disagreed that their academic performance is good.

### 3.8 CRISP-DM (Cross-Industry Standard Process for Data Mining)

The complexity of data mining has led the data analytics community to establish a standard process for data mining activities. The Cross Industry Standard Process for Data Mining (CRISP-DM) is a life cycle process for developing and analyzing data mining models (Leventhal, 2010). The CRISP-DM process is important because it gives specific tips and techniques on how to move from understanding the business data through deployment of a data mining model. CRISP-DM has six phases, which include business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Leventhal, 2010). The benefits of CRISM-DM are that it is non-proprietary and software vendor neutral, and provides a solid framework for guidance in data mining (Leventhal, 2010). The model also includes templates to aid in analysis. This process is used in a number of educational data mining studies (Luan, 2002; Vialardi et al., 2011; Y.-h. Wang & Liao, 2011). There are also a number of internal feedback loops between the phases, resulting from the very complex non-linear nature of the data mining process and ensuring the achievement of consistent and reliable results. This study is

conducted using the CRISP DM model. This model is chosen as a research methodology approach because of its application-neutral and conformity to established standards for data mining projects.

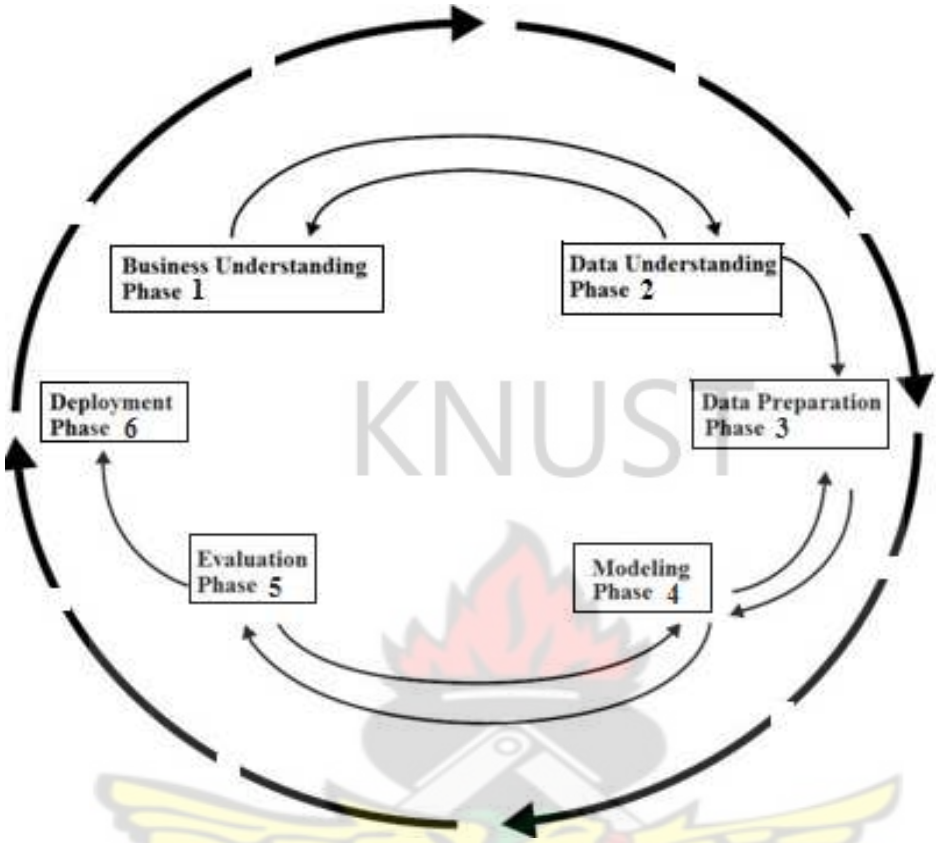


Figure 3.1 Shows the CRISP–DM is an iterative and adaptive process

The processes involved in some of the phases below have already been sufficiently dealt with by the activities in the preceding sections.

### 3.8.1 Business understanding phase

The academic performance of public basic schools in the Basic Education Certificate examination is not encouraging compared to the private basic schools. Socio-economic challenges have been identified as some of the causes of poor academic performance. According to Fraenkel and Wallen (1996), correlation research investigates the possibility of an existing relationship between variables. This study aims to investigate the correlation between these challenges and academic performance. Based on the review of related works by other researchers on these socio-economic factors, specific research questions were formulated to help achieve the projects objectives.

### 3.8.2 Data understanding phase

The dataset is made up of one pre-classified categorical (dependent) variable with five (independent) variables made up of hundred attributes each for the training set and 50 attributes each for the test set. The independent variables include parent's education level, size of family, parent's employment status,

parent's marital status and parent's income level. Some of the parameters are removed, e.g., age, gender and class are fields containing data that is of no particular interest to the research.

### 3.8.3 Data preparation phase

Having obtained the responses for the questionnaires, the process of preparing the data was accomplished. The list of collected and selected attributes which is relevant for this research is specified below.

- **FS** - Family size. The attributes of this variable is obtained through the questionnaire. It is mapped into two categories of small and large based on 100 records for the training set and 50 records for the test set. Small = 44% and large = 56% in the training set and small = 26% and large = 24% for the test set.
- **PEL** – Parents education level. The attributes of this variable is obtained through the questionnaire. It is mapped into two categories of educated and uneducated based on 100 records for the training set and 50 records for the test set. Educated = 53% and Uneducated = 47% in the training set and Educated = 31% and Uneducated = 19% in the test set.
- **PES** – Parents employment status. The attributes of this variable is obtained through the questionnaire. It is mapped into two categories of employed and unemployed based on 100 records for the training set and 50 records for the test set. Employed = 56% and unemployed = 44% in the training set and Employed = 33% and unemployed = 17% in the test set.
- **PMS** – Parents marital status. The attributes of this variable is obtained through the questionnaire. It is mapped into two categories of single and married based on 100 records for the training set and 50 records for the test set. Single = 43% and married = 57% in the training set and Single = 22% and married = 28% in the test set.
- **MI** – Monthly income. The attributes for this variable is obtained through the questionnaire. It is mapped into two categories of low and high based on 100 records for the training set and 50 records for the test set. Low = 55% and high = 45% in the training set and Low = 25% and high = 25% in the test set.
- **AP**- Academic performance. This is the target or output class. The attributes for the academic performance variable is also obtained from the questionnaire. It is mapped into two categories of good and bad based on 100 records for the training set and 50 records for the test set. Good = 46% and bad = 54% in the training set and Good = 28% and bad = 22% in the test set.

Based on the background and related work, a data set with the attributes depicted above are selected to perform the analysis on the socio-economic variables and their effect on academic performance for both the training set and test set.

### 3.8.4 Modeling phase

A decision tree model and decision rules would be generated from both the training set and the test set using the C4.5 algorithm to test the research hypotheses

### 3.8.5 Evaluation

The holdout method is used in determining the accuracy of the model. In this method, the responses from the questionnaire is partitioned into two independent sets, a training set and a test set, Han et al, (2006) . Two thirds of the data is allocated to the training set, and the remaining one third is allocated to the test set. The training set is used to derive the model, whose accuracy is estimated with the test set, Han et al, (2006). To ensure that this study is generalizable to the population, the attributes of the training set would be compared to those of the attributes of the test set to find out if the margin of difference is large or small.

### 3.8.6 Deployment Phase

The deployment is at the discretion of the schools. However, it is my belief that, this research will enable public basic schools address the effect of socio-economic challenges students face in pursuit of academic excellence.

## 3.9 Ethical Consideration

The conducting of research requires not only expertise and diligence, but also honesty and integrity. This is done to recognize and protect the rights of human subjects. To render the study ethical, the rights to self-determination, anonymity, confidentiality and informed consent were observed. Verbal permission was obtained from head teachers in charge of the selected public basic schools. Students' consent was obtained before they completed the questionnaires. Burns & Grove (1993) define informed consent as the prospective subject's agreement to participate voluntarily in a study, which is reached after assimilation of essential information about the study. The subjects were informed of their rights to voluntarily consent or decline to participate, and to withdraw participation at any time without penalty. Respondents were also assured that the study was strictly academic and that utmost confidentiality would be observed. Students were also informed about the procedures that would be used to collect the data, and were assured that there were no potential risks or costs involved. Anonymity was maintained throughout the study. The data used in this study was anonymously coded and cannot therefore be

traced back to individual students. Burns and Grove (1993) define anonymity as when subjects cannot be linked, even by the researcher, with his or her individual responses. According to Polit & Hungler (1995), when subjects are promised confidentiality it means that the information they provide will not be publicly reported in a way which identifies them. In this study, confidentiality was maintained by keeping the collected data confidential and not revealing the subjects' identities when reporting or publishing the study (Burns & Grove, 1993). No identifying information was entered onto the questionnaires, and questionnaires were only numbered after data was collected (Polit & Hungler, 1995).

# CHAPTER FOUR

# MODEL DEVELOPMENT AND EVALUATION PROCESSES

## 4.0 Introduction

This chapter presents the data obtained from the research questions regarding the topic. The development was undertaken as per the study variables. From the likert scale analysis, 100% was the response rate and was representative enough for reliability and to undertake the data mining analysis. The C4.5 data mining algorithm is used for the classification task. It is an algorithm developed by John Ross Quinlan that generates decision trees, which can be used to analyze the socio-economic variables selected for this study (Larose, 2005). The algorithm analyzes the training data set, builds a decision tree and generates decision rules. The root nodes are the top node of the tree and it considers all samples and selects the attributes that are most significant. The sample information is passed to branch nodes which eventually terminate in leaf nodes that give decisions. Rules are generated by illustrating the path from the root node to leaf node. Building this classification model, the primary goal is to make the model most accurately predict the desired target value for new data

## 4.1 Information Theory

Information theory is the branch of mathematics that describes how uncertainty should be quantified, manipulated and represented. Ever since the fundamental premises of information theory were laid down by Claude Shannon in 1949, it has had far reaching implications for almost every field of science and technology. Information theory has also had an important role in shaping theories of perception, cognition, and neural computation. The most fundamental quantity in information theory is entropy (Shannon and Weaver, 1949). Entropy: "A measure used to determine the disorder in the population," according to Shannon Information Theory (Shannon & Weaver, 1949). Shannon's formula for calculating entropy is; Entropy = - (Σ pi log2 pi ) Where log2 pi = (log 10 pi / log 10 2). Shannon borrowed the concept of entropy from thermodynamics where it describes the amount of disorder of a system. In information theory, entropy measures the amount of uncertainty of an unknown or random quantity. In engineering applications, information is analogous to signal, and entropy is analogous to noise Larose (2005)

### 4.1.1 Entropy and Information Gain.

Entropy is the measure which provides the average amount of uncertainty associated with a set of probabilities. In other words, it is simply the average (expected) amount of the information from an event. It is used in C4.5 algorithms to determine how disordered the attributes are in the data set.

Entropy is related to information in the sense that the higher the entropy or uncertainty, the more information is required in order to completely describe the data. According to Larose (2005), the C4.5 algorithm uses the concept of information gain or entropy reduction to select the optimal split. To practically illustrate the concept of entropy, an example of calculating the Information of a coin toss is used. There are two probabilities in fair coin toss, which are head(.5) and tail(.5). So if we get either head or tail, we will get 1 bit of information through following the formula below. I(head) = -log (.5) = 1 bit

**Information Equation**

$$I(p) = -\log_b(p) \qquad \text{equation 1}$$

Where;

p = probability of the event happening

b = base (base 2 is mostly used in information theory)

For variables with several outcomes, we simply use a weighted sum of the log2(pj )'s, with weights equal to the outcome probabilities, resulting in the formula

$$H(X) = -\sum_j p_j \log_2(p_j) \qquad \text{equation 2}$$

Where;

H (X) = The entropy H of a discreet random variables (one which may take on only a countable number of distinct values such as 0,1,2,3,4. eg, the number of children in a family) X (training data set)

$\log_2$ = Is the log function to the base 2 and it is used because the information is encoded in bits

j = 1, ……. ,k

pj = The probability that an arbitrary set of training data belongs to; for example, the small or large family attribute.

Suppose we have a candidate split S, (eg. size of family) which partitions the training data set T into several subsets, T1, T2, . . . , Tk. (eg. small or large family size). The mean information requirement is calculated as the weighted sum of the entropies for the individual subsets as follows;

$$H_S(T) = \sum_{i=1}^{k} P_i H_S(T_i) \qquad \text{equation 3}$$

Where;

H s(T) = Entropy H of gain S of partitioning candidate split in the training data into subsets. (eg. small or large family size).

S = Candidate split (eg. Size of family)

T = Partitioned training data into subsets of T1, T2,……Tk (eg. Small or large family size)

Pi = represents the proportion of records in subset i. (eg. Small or large family size)

Ti = Partitions training data in subset i. (eg. Small or large family size)

Hs = Entropy H of partioning candidate split. (eg. Size of family)

i = 1,……,k



The entropies of the two subsets and the proportions of the subset Pi are combined using equation 3. Information gain is utilized by the C4.5 algorithm as a measure of the effectiveness of an attribute in classifying the training data. Information gain is computed by measuring the difference between the entropy of the data set before the split H(T) and the overall entropy of the data set after the split H s(T). At each decision node, C4.5 chooses the optimal split to be the split that has the greatest information gain, gain(S). Information gain is defined as;

$$gain(S) = H(T) - H_S(T) \qquad \text{equation 4}$$

That is, the increase in information produced by partitioning the training data T (eg. Small or large family size) according to this candidate split S. (eg. Size of family)

Where;

H(T) = Entropy H of partitioned dependant variable. (eg. academic performance class into good or bad)

H s(T) = Entropy H of gain S of partitioning candidate split in the training data into subsets. (eg. small or large family size).

This is how gain (s) will be interpreted. First, H(T ) = (eg. 0.9999) will mean that, on average, one would need 0.9999 bit (0's or 1's) to transmit the (eg. Academic performance class) of the hundred students in the data set. The amount of bit generated from the computation of Hs(T) will indicated that, the partitioning of the students into two subsets will lower the average bit requirement for transmitting the academic performance status of the students. This often results in lower entropy and this entropy reduction can be viewed as information gain. This gain will be compared to the information gained by the other candidate splits, and choose the split with the largest information gain as the optimal split for the decision node.

### 4.2 Data modeling (classification)

Classification is a two-step process. In the first step, a classifier is built describing a predetermined set of data classes, (Han & Kamber, 2006). This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or learning from a training set made up of database tuples and their associated class labels. Each tuple is assumed to belong to a predetermined class as determined by another attribute called the class label attribute. It is categorical in that each value serves as a category or class. The individual tuples making up the training set are referred to as training tuples. In the context of classification, data tuples can be referred to as samples or data points, (Han & Kamber, 2006). Since the class label of each training tuple is provided, this step is also known as supervised learning (i.e., the learning of the classifier is supervised" in that it is told to which class each training tuple belongs). In the second step, a test set is used, made up of test tuples and their associated class labels. They are independent of the training tuples, meaning that they are not used to construct the classifier, (Han & Kamber, 2006).The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier, (Han & Kamber, 2006).The associated class label of each test tuple is compared with the learned classifier's class prediction for that tuple. If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known, (Han & Kamber, 2006). The modeling phase of the CRISP-DM begins with the candidate split at the root node..

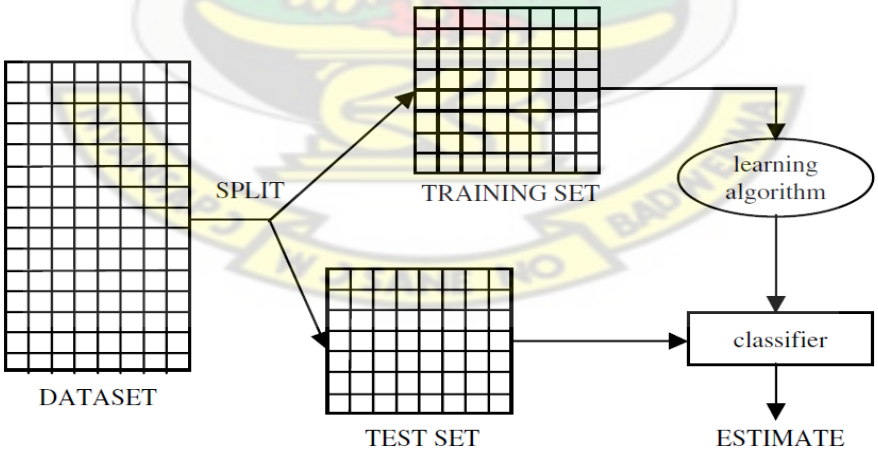


Figure 4.1 Data modeling process

#### 4.2.1 Candidate split at root node

Determining the optimal split for the root node containing records (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69,

70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100) as indicated in Table 4.1 below

Table 4.1 Training set of records at the root node

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| 1 | Small | Educated | Employed | Married | High | Good |
| 2 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 3 | Large | Uneducated | Employed | Married | High | Bad |
| 4 | Large | Educated | Unemployed | Single | Low | Good |
| 5 | Small | Uneducated | Employed | Married | High | Bad |
| 6 | Large | Educated | Employed | Married | Low | Good |
| 7 | Small | Uneducated | Employed | Married | Low | Good |
| 8 | Large | Educated | Unemployed | Single | High | Bad |
| 9 | Large | Educated | Employed | Married | Low | Good |
| 10 | Large | Educated | Employed | Single | High | Bad |
| 11 | Small | Uneducated | Unemployed | Married | Low | Good |
| 12 | Large | Educated | Employed | Single | High | Bad |
| 13 | Small | Educated | Unemployed | Married | High | Bad |
| 14 | Large | Uneducated | Employed | Single | Low | Bad |
| 15 | Small | Educated | Unemployed | Married | High | Good |
| 16 | Large | Uneducated | Employed | Married | Low | Good |
| 17 | Large | Educated | Unemployed | Single | High | Bad |
| 18 | Small | Uneducated | Employed | Married | Low | Bad |
| 19 | Large | Educated | Employed | Single | High | Bad |
| 20 | Small | Uneducated | Unemployed | Married | High | Good |
| 21 | Large | Uneducated | Employed | Married | Low | Good |
| 22 | Large | Educated | Unemployed | Single | Low | Bad |
| 23 | Small | Uneducated | Employed | Single | High | Good |
| 24 | Small | Educated | Employed | Married | Low | Bad |
| 25 | Large | Uneducated | Unemployed | Single | High | Good |
| 26 | Small | Educated | Employed | Married | High | Good |
| 27 | Large | Educated | Unemployed | Married | Low | Good |
| 28 | Small | Uneducated | Employed | Single | High | Bad |
| 29 | Large | Educated | Employed | Married | Low | Bad |
| 30 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 31 | Small | Educated | Employed | Married | High | Good |
| 32 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 33 | Small | Educated | Employed | Married | High | Bad |
| 34 | Large | Educated | Employed | Single | Low | Good |
| 35 | Small | Uneducated | Unemployed | Married | High | Bad |
| 36 | Large | Educated | Unemployed | Single | Low | Good |
| 37 | Small | Educated | Employed | Married | High | Bad |
| 38 | Small | Uneducated | Unemployed | Single | Low | Good |
| 39 | Large | Uneducated | Employed | Married | High | Bad |
| 40 | Large | Educated | Employed | Single | High | Good |
| 41 | Large | Uneducated | Unemployed | Married | Low | Good |
| 42 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 43 | Small | Educated | Employed | Married | High | Good |
| 44 | Large | Educated | Employed | Single | Low | Bad |
| 45 | Large | Uneducated | Unemployed | Married | Low | Good |
| 46 | Small | Educated | Employed | Single | High | Bad |
| 47 | Small | Uneducated | Unemployed | Married | High | Good |
| 48 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 49 | Large | Educated | Employed | Married | Low | Good |
| 50 | Small | Uneducated | Unemployed | Single | High | Bad |

| | | | | | | |
|---|---|---|---|---|---|---|
| 51 | Large | Educated | Employed | Married | Low | Good |
| 52 | Small | Educated | Unemployed | Single | High | Good |
| 53 | Large | Uneducated | Employed | Married | Low | Bad |
| 54 | Large | Educated | Employed | Single | High | Good |
| 55 | Small | Educated | Unemployed | Married | Low | Bad |
| 56 | Large | Uneducated | Employed | Married | High | Good |
| 57 | Small | Uneducated | Employed | Married | Low | Bad |
| 58 | Large | Educated | Unemployed | Married | Low | Good |
| 59 | Large | Uneducated | Employed | Married | High | Good |
| 60 | Small | Educated | Employed | Single | High | Bad |
| 61 | Large | Uneducated | Employed | Single | Low | Bad |
| 62 | Large | Educated | Unemployed | Married | High | Bad |
| 63 | Small | Uneducated | Employed | Single | Low | Good |
| 64 | Large | Educated | Unemployed | Married | High | Bad |
| 65 | Large | Uneducated | Employed | Married | Low | Good |
| 66 | Small | Educated | Unemployed | Single | Low | Bad |
| 67 | Small | Uneducated | Employed | Married | High | Good |
| 68 | Large | Educated | Unemployed | Single | Low | Bad |
| 69 | Small | Uneducated | Employed | Married | High | Good |
| 70 | Large | Educated | Unemployed | Married | Low | Bad |
| 71 | Large | Educated | Employed | Single | High | Bad |
| 72 | Large | Uneducated | Unemployed | Married | Low | Good |
| 73 | Large | Educated | Employed | Single | Low | Bad |
| 74 | Small | Uneducated | Unemployed | Married | High | Good |
| 75 | Large | Educated | Employed | Single | Low | Bad |
| 76 | Small | Uneducated | Employed | Married | Low | Bad |
| 77 | Small | Uneducated | Unemployed | Married | High | Good |
| 78 | Large | Educated | Unemployed | Single | Low | Bad |
| 79 | Large | Uneducated | Employed | Married | High | Bad |
| 80 | Large | Educated | Employed | Single | Low | Good |
| 81 | Small | Educated | Unemployed | Married | High | Bad |
| 82 | Large | Uneducated | Employed | Married | Low | Good |
| 83 | Small | Uneducated | Unemployed | Single | High | Bad |
| 84 | Large | Educated | Employed | Married | Low | Bad |
| 85 | Small | Uneducated | Unemployed | Married | High | Good |
| 86 | Small | Educated | Employed | Single | Low | Bad |
| 87 | Large | Uneducated | Unemployed | Married | High | Good |
| 88 | Small | Educated | Employed | Single | Low | Bad |
| 89 | Small | Uneducated | Unemployed | Married | Low | Bad |
| 90 | Large | Educated | Employed | Single | Low | Good |
| 91 | Small | Uneducated | Unemployed | Married | High | Bad |
| 92 | Small | Uneducated | Employed | Married | Low | Good |
| 93 | Large | Educated | Unemployed | Single | Low | Good |
| 94 | Small | Uneducated | Employed | Married | Low | Bad |
| 95 | Large | Educated | Unemployed | Single | High | Bad |
| 96 | Small | Educated | Employed | Married | Low | Good |
| 97 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 98 | Small | Educated | Employed | Married | High | Good |
| 99 | Large | Educated | Unemployed | Single | Low | Bad |
| 100 | Small | Uneducated | Employed | Married | Low | Bad |

Entropy before splitting

46 of the 100 records are classified as good academic performance, with the remaining 54 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{46}{100}, \text{PBAP} = \frac{54}{100}$$

$$H(T) = -\sum_{j} P_j \log_2 (P_j)$$

$$-\frac{46}{100}\log_2\left(\frac{46}{100}\right)-\frac{54}{100}\log_2\left(\frac{54}{100}\right)$$

$$=-0.46\log_2(0.46)-0.54\log_2(0.54))$$

$$=-0.46\,(-1.120)-0.54\,(0.888)=0.994\text{ bit}$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.994 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 44 records have 'Small family' size and 56 records have 'Large family' size.

$$P_{small}=\frac{44}{100},\ P_{large}=\frac{56}{100}$$

Entropy for small family size

44 records (21 good, 23 bad)

$$-\frac{21}{44}\log_2\left(\frac{21}{44}\right)-\frac{23}{44}\log_2\left(\frac{23}{44}\right)$$

$$=-0.477\log_2(0.477)-0.523\log_2(0.523)$$

$$=-0.477\,(-1.067)-0.523\,(-0.935)=0.997$$

Entropy for large family size

7 records (4 good, 3 bad)

$$-\frac{25}{56}\log_2\left(\frac{25}{56}\right)-\frac{31}{56}\log_2\left(\frac{31}{56}\right)$$

$$=-0.446\log_2(0.446)-0.554\log_2(0.554)$$

$$=-0.446\,(-1.164)-0.554\,(-0.852)=0.991$$

Combine the entropies of these two subsets

$$\frac{44}{100}(0.997)+\frac{56}{100}(0.991)$$

$$=0.44\,(0.997)+0.56\,(0.991)=0.993$$

The information gain represented by the split on the size of family attribute is;

$$\text{gain}(S)=H(T)-H_S(T)$$

0.994 - 0.993 = **0.001 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), 53 records have 'Educated' status and 47 records have 'Uneducated' status.

$$P_{educated} = \frac{53}{100}, P_{uneducated} = \frac{47}{100}$$

Entropy for educated status

53 records (23 good, 30 bad)

$$-\frac{23}{53}\log_2\left(\frac{23}{53}\right) - \frac{30}{53}\log_2\left(\frac{30}{53}\right)$$

$$= -0.434\log_2(0.434) - 0.566\log_2(0.566)$$

$$= -0.434(-1.204) - 0.566(-0.821) = 0.987$$

Entropy for uneducated status

47 records (23 good, 24 bad)

$$-\frac{23}{47}\log_2\left(\frac{23}{47}\right) - \frac{24}{47}\log_2\left(\frac{24}{47}\right)$$

$$= -0.489\log_2(0.489) - 0.511\log_2(0.511)$$

$$= -0.489(-1.032) - 0.511(-0.968) = 0.999$$

Combine the entropies of these two subsets,

$$\frac{53}{100}(0.987) + \frac{47}{100}(0.999)$$

$$= 0.53(0.987) + 0.47(0.999) = 0.992$$

The information gain represented by the split on the Parents education level attribute is;

$$gain(S) = H(T) - H_S(T)$$

0.994 - 0.992 = **0.002 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), 44 records have 'Unemployed' status and 56 records have 'Employed' status.

$$P_{unemployed} = \frac{44}{100}, P_{employed} = \frac{56}{100}$$

Entropy for unemployed status

44 records (19 good, 25 bad)

$$-\frac{19}{44}\log_2\left(\frac{19}{44}\right) - \frac{25}{44}\log_2\left(\frac{25}{44}\right)$$

$$= -0.432\log_2(0.432) - 0.568\log_2(0.568)$$

$$= -0.432(-1.211) - 0.568(-0.816) = 0.986$$

Entropy for employed status

56 records (27 good, 28 bad)

$$-\frac{27}{56}\log_2\left(\frac{27}{56}\right) - \frac{28}{56}\log_2\left(\frac{28}{56}\right)$$

$$= -0.482\log_2(0.482) - 0.5\log_2(0.5)$$

$$= -0.482(-1.052) - 0.5(-1) = 1.00$$

Combine the entropies of these two subsets,

$$\frac{44}{100}(0.986) + \frac{56}{100}(1.00)$$

$$= 0.44(0.986) + 0.56(1.00) = 0.993$$

The information gain represented by the split on the Parents employment status attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

$0.994 - 0.993 =$ **0.001 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), 57 records have 'Married' status and 43 records have 'Single' status.

$$P_{\text{married}} = \frac{57}{100},\ P_{\text{single}} = \frac{43}{43}$$

Entropy for married status

57 records (33 good, 24 bad)

$$-\frac{33}{57}\log_2\left(\frac{33}{57}\right) - \frac{24}{57}\log_2\left(\frac{24}{57}\right)$$

$$= -0.579\log_2(0.579) - 0.421\log_2(0.421)$$

$$= -0.579(-0.788) - 0.421(-1.248) = 0.98$$

Entropy for single status

43 records (13 good, 30 bad)

$$-\frac{13}{43}\log_2\left(\frac{13}{43}\right) - \frac{30}{43}\log_2\left(\frac{30}{43}\right)$$

$$= -0.302\log_2(0.302) - 0.698\log_2(0.698)$$

$$= -0.302(-1.727) - 0.698(-0.518) = 0.883$$

Combine the entropies of these two subsets,

$$\frac{57}{100}(0.98) + \frac{43}{100}(0.883)$$

$$= 0.57(0.98) + 0.43(0.883) = 0.938$$

The information gain represented by the split on the Parent marital status attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.994 – 0.938 = **0.056 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 55 records have low income and 45 records have high income.

$$P_{\text{low income}} = \frac{55}{100},\ P_{\text{high income}} = \frac{45}{100}$$

Entropy for low income

55 records (25 good, 30 bad)

$$-\frac{25}{55}\log_2\left(\frac{25}{55}\right) - \frac{30}{55}\log_2\left(\frac{30}{55}\right)$$

$$= -0.455\log_2(0.455) - 0.545\log_2(0.545)$$

$$= -0.455\,(-1.136) - 0.545\,(-0.875) = 0.993$$

Entropy for high income

45 records (21 good, 24 bad)

$$-\frac{21}{45}\log_2\left(\frac{21}{45}\right) - \frac{24}{45}\log_2\left(\frac{24}{45}\right)$$

$$= -0.467\log_2(0.467) - 0.533\log_2(0.533)$$

$$= -0.467\,(-1.098) - 0.533\,(-0.907) = 0.996$$

Combine the entropies of these two subsets,

$$\frac{55}{100}(0.993) + \frac{45}{100}(0.996)$$

$$= 0.45\,(0.996) + 0.55\,(0.993) = 0.994$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.994 – 0.994 = **0.000 bit**

Table 4.2 summarizes the information gain for each candidate split at the root node. Candidate split 4, Parents marital status has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.1

Table 4.2 Information Gain for each candidate split at the root node

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.001 bits |
| 2 | Parent's education level = Educated<br>Parent's education level = Uneducated | 0.002 bits |
| 3 | Parent's employment status = Employed<br>Parent's employment status = Unemployed | 0.001 bits |
| 4 | Parents marital status = Single<br>Parents marital status = Married | 0.056 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.000 bits |




Figure 4.2 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of two new decision nodes. The 43 records at decision node 1 (Parents marital status = single) contain both good and bad academic performance. In the same vein, the 57 records at decision node 2 (Parents marital status = married) also contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.2 Candidate split at decision node 1

Determining the optimal split for decision node 1, containing records (2, 4, 8, 10, 12, 14, 17, 19, 22, 23, 25, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 60, 61, 63, 66, 68, 71, 73, 75, 78, 80, 83, 86, 88, 90, 93, 95, 97, 99) as indicated in Table 4.3 below.

Table 4.3 Training set of records available at decision node 1

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| 2 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 4 | Large | Educated | Unemployed | Single | Low | Good |
| 8 | Large | Educated | Unemployed | Single | High | Bad |
| 10 | Large | Educated | Employed | Single | High | Bad |
| 12 | Large | Educated | Employed | Single | High | Bad |
| 14 | Large | Uneducated | Employed | Single | Low | Bad |
| 17 | Large | Educated | Unemployed | Single | High | Bad |
| 19 | Large | Educated | Employed | Single | High | Bad |
| 22 | Large | Educated | Unemployed | Single | Low | Bad |
| 23 | Small | Uneducated | Employed | Single | High | Good |
| 25 | Large | Uneducated | Unemployed | Single | High | Good |
| 28 | Small | Uneducated | Employed | Single | High | Bad |
| 30 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 32 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 34 | Large | Educated | Employed | Single | Low | Good |
| 36 | Large | Educated | Unemployed | Single | Low | Good |
| 38 | Small | Uneducated | Unemployed | Single | Low | Good |
| 40 | Large | Educated | Employed | Single | High | Good |
| 42 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 44 | Large | Educated | Employed | Single | Low | Bad |
| 46 | Small | Educated | Employed | Single | High | Bad |
| 48 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 50 | Small | Uneducated | Unemployed | Single | High | Bad |
| 52 | Small | Educated | Unemployed | Single | High | Good |
| 54 | Large | Educated | Employed | Single | High | Good |
| 60 | Small | Educated | Employed | Single | High | Bad |
| 61 | Large | Uneducated | Employed | Single | Low | Bad |
| 63 | Small | Uneducated | Employed | Single | Low | Good |
| 66 | Small | Educated | Unemployed | Single | Low | Bad |
| 68 | Large | Educated | Unemployed | Single | Low | Bad |
| 71 | Large | Educated | Employed | Single | High | Bad |
| 73 | Large | Educated | Employed | Single | Low | Bad |
| 75 | Large | Educated | Employed | Single | Low | Bad |
| 78 | Large | Educated | Unemployed | Single | Low | Bad |
| 80 | Large | Educated | Employed | Single | Low | Good |
| 83 | Small | Uneducated | Unemployed | Single | High | Bad |
| 86 | Small | Educated | Employed | Single | Low | Bad |
| 88 | Small | Educated | Employed | Single | Low | Bad |
| 90 | Large | Educated | Employed | Single | Low | Good |
| 93 | Large | Educated | Unemployed | Single | Low | Good |
| 95 | Large | Educated | Unemployed | Single | High | Bad |
| 97 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 99 | Large | Educated | Unemployed | Single | Low | Bad |

Entropy before splitting

13 of the 43 records are classified as good academic performance, with the remaining 30 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{13}{43}, \text{PBAP} = \frac{30}{43}$$

$$H(T) = -\sum_{j} P\,j \log_2 (P\,j)$$

$$-\frac{13}{43}\log_2\left(\frac{13}{43}\right) - \frac{30}{43}\log_2\left(\frac{30}{43}\right)$$

$$= -0.302 \log_2(0.302) - 0.698 \log_2(0.698)$$

$$= -0.302\,(-1.727) - 0.698\,(-0.518) = 0.883$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.883 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 12 records have 'Small family' size and 31 records have 'Large family' size.

$$P_{small} = \frac{12}{43},\ P_{large} = \frac{31}{43}$$

Entropy for small family size

12 records (4 good, 8 bad)

$$-\frac{4}{12}\log_2\left(\frac{4}{12}\right) - \frac{8}{12}\log_2\left(\frac{8}{12}\right)$$

$$= -0.333\log_2(0.333) - 0.667\log_2(0.667)$$

$$= -0.333\,(-1.586) - 0.667\,(-0.584) = 0.917$$

Entropy for large family size

31 records (9 good, 22 bad)

$$-\frac{9}{31}\log_2\left(\frac{9}{31}\right) - \frac{22}{31}\log_2\left(\frac{22}{31}\right)$$

$$= -0.29\log_2(0.29) - 0.71\log_2(0.71)$$

$$= -0.29\,(-1.785) - 0.71\,(-0.494) = 0.868$$

Combine the entropies of these two subsets

$$\frac{12}{43}(0.917) + \frac{31}{43}(0.868)$$

$$= 0.279\,(0.917) + 0.721\,(0.868) = 0.881$$

The information gain represented by the split on the size of family attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.883 - 0.881 = **0.002 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), 28 records have 'Educated' status and 15 records have 'Uneducated' status.

$$P_{educated} = \frac{28}{43},\ P_{uneducated} = \frac{15}{43}$$

Entropy for educated status

28 records (9 good, 19 bad)

$$-\frac{9}{28}\log_2\left(\frac{9}{28}\right)-\frac{19}{28}\log_2\left(\frac{19}{28}\right)$$

$$=-0.321\log_2(0.321)-0.679\log_2(0.679)$$

$$=-0.321\,(-1.639)-0.679\,(-0.558)=0.905$$

Entropy for uneducated status

15 records (4 good, 11 bad)

$$-\frac{4}{15}\log_2\left(\frac{4}{15}\right)-\frac{11}{15}\log_2\left(\frac{11}{15}\right)$$

$$=-0.267\log_2(0.267)-0.733\log_2(0.733)$$

$$=-0.267\,(-1.905)-0.733\,(-0.448)=0.837$$

Combine the entropies of these two subsets,

$$\frac{28}{43}(0.905)+\frac{15}{43}(0.837)$$

$$=0.651(0.905)+0.349\,(0.837)=0.881$$

The information gain represented by the split on the Parents education level attribute is;

$$\text{gain}(S)=H(T)-H_S(T)$$

0.883 - 0.881 = **0.002 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), 22 records have 'Unemployed' status and 21 records have 'Employed' status.

$$P_{\text{unemployed}}=\frac{22}{43},\ P_{\text{employed}}=\frac{21}{43}$$

Entropy for unemployed status

22 records (6 good, 16 bad)

$$-\frac{19}{44}\log_2\left(\frac{19}{44}\right)-\frac{25}{44}\log_2\left(\frac{25}{44}\right)$$

$$=-0.273\log_2(0.273)-0.727\log_2(0.727)$$

$$=-0.273\,(-1.873)-0.727\,(-0.459)=0.845$$

Entropy for employed status

21 records (7 good, 14 bad)

$$-\frac{7}{21}\log_2\left(\frac{7}{21}\right)-\frac{14}{21}\log_2\left(\frac{14}{21}\right)$$

$$=-0.333\log_2(0.333)-0.667\log_2(0.667)$$

$$= -0.333\,(-1.586) - 0.667\,(-0.584) = 0.917$$

Combine the entropies of these two subsets,

$$\frac{22}{43}(0.845) + \frac{21}{43}(0.917)$$

$$= 0.512\,(0.845) + 0.488\,(0.917) = 0.88$$

The information gain represented by the split on the Parents employment status attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.883 – 0.880 = **0.003 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 43 records have 'Single' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 26 records have low income and 17 records have high income.

$$P_{\text{low income}} = \frac{26}{43},\ P_{\text{high income}} = \frac{17}{43}$$

Entropy for low income

26 records (8 good, 18 bad)

$$-\frac{8}{26}\log_2\left(\frac{8}{26}\right) - \frac{18}{26}\log_2\left(\frac{18}{26}\right)$$

$$= -0.308\log_2(0.308) - 0.692\log_2(0.692)$$

$$= -0.308\,(-1.698) - 0.692\,(-0.531) = 0.890$$

Entropy for high income

17 records (5 good, 12 bad)

$$-\frac{5}{17}\log_2\left(\frac{5}{17}\right) - \frac{12}{17}\log_2\left(\frac{12}{17}\right)$$

$$= -0.294\log_2(0.294) - 0.706\log_2(0.706)$$

$$= -0.294\,(-1.766) - 0.706\,(-0.502) = 0.873$$

Combine the entropies of these two subsets,

$$\frac{26}{43}(0.890) + \frac{17}{43}(0.873)$$

$$= 0.605\,(0.890) + 0.395\,(0.873) = 0.883$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

$$0.883 - 0.883 = \textbf{0.000 bit}$$

Table 4.4 summarizes the information gain for each candidate split at decision node 1. Candidate split 3, Parents employment status has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.3

Table 4.4 Information Gain for each candidate split at decision node 1

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.002 bits |
| 2 | Parent's education level = Educated<br>Parent's education level = Uneducated | 0.002 bits |
| 3 | Parent's employment status = Employed<br>Parent's employment status = Unemployed | 0.003 bits |
| 4 | Parents marital status = Single | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.000 bits |

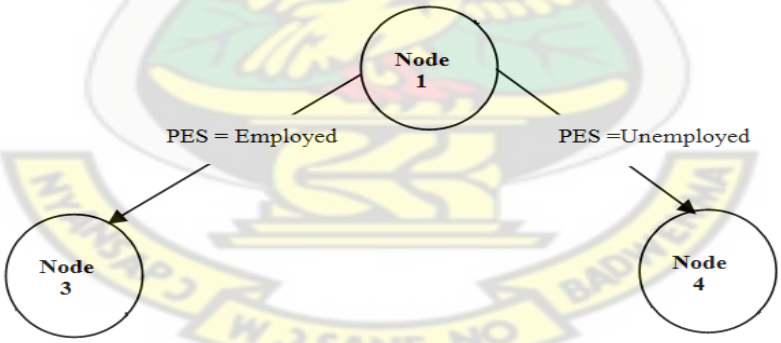


Figure 4.3 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of two new decision nodes. The 21 records at decision node 3 (Parents employment status = employed) contain both good and bad academic performance. In the same vein, the 22 records at decision node 4 (Parents employment status = unemployed) also contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.3 Candidate split at decision node 2

Determining the optimal split for decision node 2, containing records (1, 3, 5, 6, 7, 9, 11, 13, 15, 16, 18, 20, 21, 24, 26, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 56, 57, 58, 59, 62, 64, 65, 67, 69,

70, 72, 74, 76, 77, 79, 81, 82, 84, 85, 87, 89, 91, 92, 94, 96, 98, 100) as indicated in Table 4.5 below.

Table 4.5 Training set of records available at decision node 2

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| 1 | Small | Educated | Employed | Married | High | Good |
| 3 | Large | Uneducated | Employed | Married | High | Bad |
| 5 | Small | Uneducated | Employed | Married | High | Bad |
| 6 | Large | Educated | Employed | Married | Low | Good |
| 7 | Small | Uneducated | Employed | Married | Low | Good |
| 9 | Large | Educated | Employed | Married | Low | Good |
| 11 | Small | Uneducated | Unemployed | Married | Low | Good |
| 13 | Small | Educated | Unemployed | Married | High | Bad |
| 15 | Small | Educated | Unemployed | Married | High | Good |
| 16 | Large | Uneducated | Employed | Married | Low | Good |
| 18 | Small | Uneducated | Employed | Married | Low | Bad |
| 20 | Small | Uneducated | Unemployed | Married | High | Good |
| 21 | Large | Uneducated | Employed | Married | Low | Good |
| 24 | Small | Educated | Employed | Married | Low | Bad |
| 26 | Small | Educated | Employed | Married | High | Good |
| 27 | Large | Educated | Unemployed | Married | Low | Good |
| 29 | Large | Educated | Employed | Married | Low | Bad |
| 31 | Small | Educated | Employed | Married | High | Good |
| 33 | Small | Educated | Employed | Married | High | Bad |
| 35 | Small | Uneducated | Unemployed | Married | High | Bad |
| 37 | Small | Educated | Employed | Married | High | Bad |
| 39 | Large | Uneducated | Employed | Married | High | Bad |
| 41 | Large | Uneducated | Unemployed | Married | Low | Good |
| 43 | Small | Educated | Employed | Married | High | Good |
| 45 | Large | Uneducated | Unemployed | Married | Low | Good |
| 47 | Small | Uneducated | Unemployed | Married | High | Good |
| 49 | Large | Educated | Employed | Married | Low | Good |
| 51 | Large | Educated | Employed | Married | Low | Good |
| 53 | Large | Uneducated | Employed | Married | Low | Bad |
| 55 | Small | Educated | Unemployed | Married | Low | Bad |
| 56 | Large | Uneducated | Employed | Married | High | Good |
| 57 | Small | Uneducated | Employed | Married | Low | Bad |
| 58 | Large | Educated | Unemployed | Married | Low | Good |
| 59 | Large | Uneducated | Employed | Married | High | Good |
| 62 | Large | Educated | Unemployed | Married | High | Bad |
| 64 | Large | Educated | Unemployed | Married | High | Bad |
| 65 | Large | Uneducated | Employed | Married | Low | Good |
| 67 | Small | Uneducated | Employed | Married | High | Good |
| 69 | Small | Uneducated | Employed | Married | High | Good |
| 70 | Large | Educated | Unemployed | Married | Low | Bad |
| 72 | Large | Uneducated | Unemployed | Married | Low | Good |
| 74 | Small | Uneducated | Unemployed | Married | High | Good |
| 76 | Small | Uneducated | Employed | Married | Low | Bad |
| 77 | Small | Uneducated | Unemployed | Married | High | Good |
| 79 | Large | Uneducated | Employed | Married | High | Bad |
| 81 | Small | Educated | Unemployed | Married | High | Bad |
| 82 | Large | Uneducated | Employed | Married | Low | Good |
| 84 | Large | Educated | Employed | Married | Low | Bad |
| 85 | Small | Uneducated | Unemployed | Married | High | Good |
| 87 | Large | Uneducated | Unemployed | Married | High | Good |
| 89 | Small | Uneducated | Unemployed | Married | Low | Bad |
| 91 | Small | Uneducated | Unemployed | Married | High | Bad |
| 92 | Small | Uneducated | Employed | Married | Low | Good |
| 94 | Small | Uneducated | Employed | Married | Low | Bad |
| 96 | Small | Educated | Employed | Married | Low | Good |
| 98 | Small | Educated | Employed | Married | High | Good |
| 100 | Small | Uneducated | Employed | Married | Low | Bad |

Entropy before splitting

35 of the 57 records are classified as good academic performance, with the remaining 24 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{33}{57}, \text{PBAP} = \frac{24}{57}$$

$$H(T) = -\sum_j P_j \log_2 (P_j)$$

$$-\frac{33}{57}\log_2\left(\frac{33}{57}\right) - \frac{24}{57}\log_2\left(\frac{24}{57}\right)$$

$$= -0.579 \log_2(0.579) - 0.421 \log_2(0.421)$$

$$= -0.579\,(-0.788) - 0.421\,(-1.248) = 0.982 \text{ bit}$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.982 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 32 records have 'Small family' size and 25 records have 'Large family' size.

$$P_{small} = \frac{32}{57}, P_{large} = \frac{25}{57}$$

Entropy for small family size

32 records (17 good, 19 bad)

$$-\frac{17}{32}\log_2\left(\frac{17}{32}\right) - \frac{19}{32}\log_2\left(\frac{19}{32}\right)$$

$$= -0.531 \log_2(0.531) - 0.469 \log_2(0.469)$$

$$= -0.531\,(-0.913) - 0.469\,(-1.092) = 0.996$$

Entropy for large family size

25 records (16 good, 9 bad)

$$-\frac{16}{25}\log_2\left(\frac{16}{25}\right) - \frac{9}{25}\log_2\left(\frac{9}{25}\right)$$

$$= -0.64 \log_2(0.64) - 0.36 \log_2(0.36)$$

$$= -0.64\,(-0.644) - 0.36\,(-1.474) = 0.943$$

Combine the entropies of these two subsets

$$\frac{32}{57}(0.996) + \frac{25}{57}(0.943)$$

$$= 0.561\,(0.996) + 0.438\,(0.943) = 0.972$$

The information gain represented by the split on the size of family attribute is;

$$\text{gain(S)} = H(T) - H_S(T)$$

0.982 - 0.972 = **0.010 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), 24 records have 'Educated' status and 33 records have 'Uneducated' status.

$$P_{educated} = \frac{24}{57}, P_{uneducated} = \frac{33}{57}$$

Entropy for educated status

24 records (14 good, 10 bad)

$$-\frac{14}{24}\log_2\left(\frac{14}{24}\right) - \frac{10}{24}\log_2\left(\frac{10}{24}\right)$$

$$= -0.583\log_2(0.583) - 0.417\log_2(0.417)$$

$$= -0.583(-0.778) - 0.417(-1.262) = 0.979$$

Entropy for uneducated status

33 records (20 good, 13 bad)

$$-\frac{20}{33}\log_2\left(\frac{20}{33}\right) - \frac{13}{33}\log_2\left(\frac{13}{33}\right)$$

$$= -0.606\log_2(0.606) - 0.394\log_2(0.394)$$

$$= -0.606(-0.723) - 0.394(-1.344) = 0.967$$

Combine the entropies of these two subsets,

$$\frac{24}{57}(0.979) + \frac{33}{57}(0.967)$$

$$= 0.421(0.979) + 0.578(0.967) = 0.971$$

The information gain represented by the split on the Parents education level attribute is;

$$gain(S) = H(T) - H_S(T)$$

0.982 - 0.971 = **0.011 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), 22 records have 'Unemployed' status and 35 records have 'Employed' status.

$$P_{unemployed} = \frac{22}{57}, P_{employed} = \frac{35}{57}$$

Entropy for unemployed status

22 records (13 good, 9 bad)

$$-\frac{13}{22}\log_2\left(\frac{13}{22}\right) - \frac{9}{22}\log_2\left(\frac{9}{22}\right)$$

$$= -0.591\log_2(0.591) - 0.409\log_2(0.409)$$

$$= -0.591\,(-0.759) - 0.409\,(-1.289) = 0.976$$

Entropy for employed status

35 records (20 good, 15 bad)

$$-\frac{20}{35}\log_2\left(\frac{20}{35}\right) - \frac{15}{35}\log_2\left(\frac{15}{35}\right)$$

$$= -0.571\log_2(0.571) - 0.429\log_2(0.429)$$

$$= -0.571\,(-0.808) - 0.429\,(-1.221) = 0.985$$

Combine the entropies of these two subsets,

$$\frac{35}{57}(0.985) + \frac{22}{57}(0.976)$$

$$= 0.614\,(0.985) + 0.386\,(0.976) = 0.981$$

The information gain represented by the split on the Parents employment status attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.982 – 0.981 = **0.001 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 57 records have 'Married' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 29 records have low income and 28 records have high income.

$$P_{\text{low income}} = \frac{29}{57},\ P_{\text{high income}} = \frac{28}{57}$$

Entropy for low income

29 records (17 good, 12 bad)

$$-\frac{17}{29}\log_2\left(\frac{17}{29}\right) - \frac{12}{29}\log_2\left(\frac{12}{29}\right)$$

$$= -0.586\log_2(0.586) - 0.414\log_2(0.414)$$

$$= -0.586\,(-0.771) - 0.414\,(-1.272) = 0.978$$

Entropy for high income

28 records (16 good, 12 bad)

$$-\frac{16}{28}\log_2\left(\frac{16}{28}\right)-\frac{12}{28}\log_2\left(\frac{12}{28}\right)$$

$$=-0.571\log_2(0.571)-0.429\log_2(0.429)$$

$$=-0.571(-0.808)-0.429(-1.221)=0.985$$

Combine the entropies of these two subsets,

$$\frac{29}{57}(0.978)+\frac{28}{57}(0.985)$$

$$=0.509(0.978)+0.491(0.985)=0.981$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

$$0.982 - 0.981 = \mathbf{0.001\ bit}$$

Table 4.6 summarizes the information gain for each candidate split at decision node 2. Candidate split 2, Parents education level has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.4

Table 4.6 Information Gain for each candidate split at decision node 2

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.001 bits |
| 2 | Parent's education level = Educated<br>Parent's education level = Uneducated | 0.011 bits |
| 3 | Parent's employment status = Employed<br>Parent's employment status = Unemployed | 0.001 bits |
| 4 | Parents marital status = Married | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.001 bits |

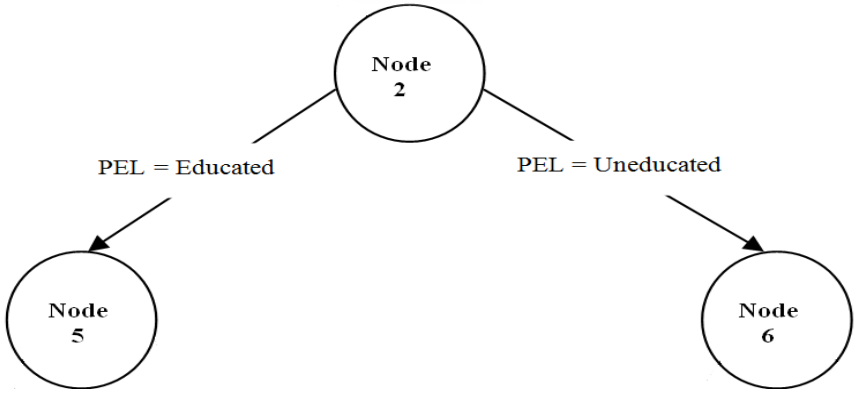


Figure 4.4 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of two new decision nodes. The 24 records at decision node 5 (Parents education level = educated) contain both good and bad academic performance. In the same vein, the 33 records at decision node 6 (Parents education level = uneducated) also contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.4 Candidate split at decision node 3

Determining the optimal split for decision node 3, containing records (10, 12, 14, 19, 23, 28, 34, 40, 44, 46, 54, 60, 61, 63, 71, 73, 75, 80, 86, 88, 90) as indicated in Table 4.7 below.

Table 4.7 Training set of records available at decision node 3

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **10** | Large | Educated | Employed | Single | High | Bad |
| **12** | Large | Educated | Employed | Single | High | Bad |
| **14** | Large | Uneducated | Employed | Single | Low | Bad |
| **19** | Large | Educated | Employed | Single | High | Bad |
| **23** | Small | Uneducated | Employed | Single | High | Good |
| **28** | Small | Uneducated | Employed | Single | High | Bad |
| **34** | Large | Educated | Employed | Single | Low | Good |
| **40** | Large | Educated | Employed | Single | High | Good |
| **44** | Large | Educated | Employed | Single | Low | Bad |
| **46** | Small | Educated | Employed | Single | High | Bad |
| **54** | Large | Educated | Employed | Single | High | Good |
| **60** | Small | Educated | Employed | Single | High | Bad |
| **61** | Large | Uneducated | Employed | Single | Low | Bad |
| **63** | Small | Uneducated | Employed | Single | Low | Good |
| **71** | Large | Educated | Employed | Single | High | Bad |
| **73** | Large | Educated | Employed | Single | Low | Bad |
| **75** | Large | Educated | Employed | Single | Low | Bad |
| **80** | Large | Educated | Employed | Single | Low | Good |
| **86** | Small | Educated | Employed | Single | Low | Bad |
| **88** | Small | Educated | Employed | Single | Low | Bad |
| **90** | Large | Educated | Employed | Single | Low | Good |

Entropy before splitting

7 of the 21 records are classified as good academic performance, with the remaining 14 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{7}{21}, \text{PBAP} = \frac{14}{21}$$

$$H(T) = -\sum_{j} P\,j \log_2 (P\,j)$$

$$-\frac{7}{21}\log_2\left(\frac{7}{21}\right) - \frac{14}{21}\log_2\left(\frac{14}{21}\right)$$

$$= -0.333 \log_2(0.333) - 0.667 \log_2(0.667)$$

$$= -0.333(-1.586) - 0.667(-0.584) = 0.917$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.917 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 7 records have 'Small family' size and 14 records have 'Large family' size.

$$P_{small} = \frac{7}{21}, P_{large} = \frac{14}{21}$$

Entropy for small family size

7 records (2 good, 5 bad)

$$-\frac{2}{7}\log_2\left(\frac{2}{7}\right) - \frac{5}{7}\log_2\left(\frac{5}{7}\right)$$

$$= -0.286\log_2(0.286) - 0.714\log_2(0.714)$$

$$= -0.286(-1.806) - 0.714(-0.486) = 0.863$$

Entropy for large family size

14 records (4 good, 10 bad)

$$-\frac{4}{14}\log_2\left(\frac{4}{14}\right) - \frac{10}{14}\log_2\left(\frac{10}{14}\right)$$

$$= -0.286\log_2(0.286) - 0.714\log_2(0.714)$$

$$= -0.286(-1.806) - 0.714(-0.486) = 0.863$$

Combine the entropies of these two subsets

$$\frac{7}{21}(0.863) + \frac{14}{21}(0.863)$$

$$= 0.333(0.863) + 0.667(0.863) = 0.863$$

The information gain represented by the split on the size of family attribute is;

$$gain(S) = H(T) - H_S(T)$$

0.917 - 0.863 = **0.054 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), 16 records have 'Educated' status and 5 records have 'Uneducated' status.

$$P_{educated} = \frac{16}{21}, P_{uneducated} = \frac{5}{21}$$

Entropy for educated status

16 records (5 good, 11 bad)

$$-\frac{5}{16}\log_2\left(\frac{5}{16}\right) - \frac{11}{16}\log_2\left(\frac{11}{16}\right)$$

$$= -0.312\log_2(0.312) - 0.688\log_2(0.688)$$

$$= -0.312\,(-1.68) - 0.688\,(-0.539) = 0.894$$

Entropy for uneducated status

5 records (2 good, 3 bad)

$$-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$= -0.4\log_2(0.4) - 0.6\log_2(0.6)$$

$$= -0.4\,(-1.321) - 0.6\,(-0.736) = 0.97$$

Combine the entropies of these two subsets,

$$\frac{16}{21}(0.894) + \frac{5}{21}(0.97)$$

$$= 0.761\,(0.894) + 0.238\,(0.97) = 0.911$$

The information gain represented by the split on the Parents education level attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.917 - 0.911 = **0.006 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), all 21 records have 'Employed' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 21 records have 'Single' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 11 records have low income and 10 records have high income.

$$P_{\text{low income}} = \frac{11}{21},\ P_{\text{high income}} = \frac{10}{21}$$

Entropy for low income

11 records (4 good, 7 bad)

$$-\frac{4}{11}\log_2\left(\frac{4}{11}\right)-\frac{7}{11}\log_2\left(\frac{7}{11}\right)$$

$$=-0.364\log_2(0.364)-0.636\log_2(0.636)$$

$$=-0.364\,(-1.457)-0.636\,(-0.653)=0.945$$

Entropy for high income

10 records (3 good, 7 bad)

$$-\frac{3}{10}\log_2\left(\frac{3}{10}\right)-\frac{7}{10}\log_2\left(\frac{7}{10}\right)$$

$$=-0.3\log_2(0.3)-0.7\log_2(0.7)$$

$$=-0.3\,(-1.736)-0.7\,(-0.514)=0.881$$

Combine the entropies of these two subsets,

$$\frac{11}{21}(0.945)+\frac{10}{21}(0.881)$$

$$=0.524\,(0.945)+0.476\,(0.881)=0.914$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S)=H(T)-H_S(T)$$

$0.917 - 0.914 =$ **0.003 bit**

Table 4.8 summarizes the information gain for each candidate split at decision node 3. Candidate split 1, Size of family has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.5

Table 4.8 Information Gain for each candidate split at decision node 3

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.054 bits |
| 2 | Parent's education level = Educated<br>Parent's education level = Uneducated | 0.006 bits |
| 3 | Parent's employment status = Employed | 0.000 bits |
| 4 | Parents marital status = Single | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.003 bits |

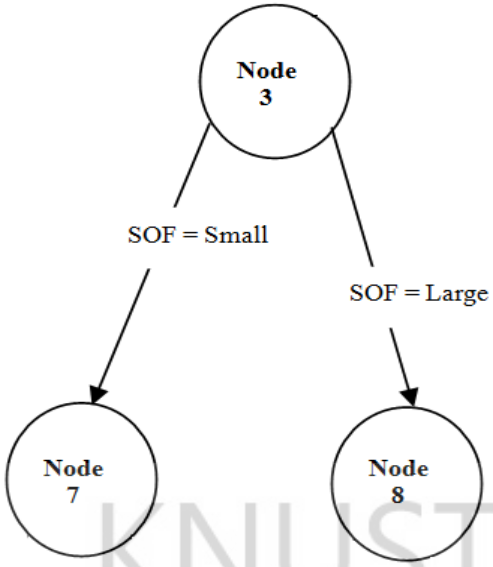


Figure 4.5 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of two new decision nodes. The 7 records at decision node 7 (Size of family = small) contain both good and bad academic performance. In the same vein, the 14 records at decision node 8 (Size of family = large) also contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.5 Candidate split at decision node 4

Determining the optimal split for decision node 4, containing records (2, 4, 8, 17, 22, 23, 25, 30, 32, 36, 38, 42, 48, 50, 52, 66, 68, 78, 83, 93, 95, 97, 99) as indicated in Table 4.9 below.

Table 4.9 Training set of records available at decision node 4

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| 2 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 4 | Large | Educated | Unemployed | Single | Low | Good |
| 8 | Large | Educated | Unemployed | Single | High | Bad |
| 17 | Large | Educated | Unemployed | Single | High | Bad |
| 22 | Large | Educated | Unemployed | Single | Low | Bad |
| 25 | Large | Uneducated | Unemployed | Single | High | Good |
| 30 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 32 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 36 | Large | Educated | Unemployed | Single | Low | Good |
| 38 | Small | Uneducated | Unemployed | Single | Low | Good |
| 42 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 48 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 50 | Small | Uneducated | Unemployed | Single | High | Bad |
| 52 | Small | Educated | Unemployed | Single | High | Good |
| 66 | Small | Educated | Unemployed | Single | Low | Bad |
| 68 | Large | Educated | Unemployed | Single | Low | Bad |
| 78 | Large | Educated | Unemployed | Single | Low | Bad |
| 83 | Small | Uneducated | Unemployed | Single | High | Bad |
| 93 | Large | Educated | Unemployed | Single | Low | Good |
| 95 | Large | Educated | Unemployed | Single | High | Bad |
| 97 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 99 | Large | Educated | Unemployed | Single | Low | Bad |

Entropy before splitting

6 of the 22 records are classified as good academic performance, with the remaining 16 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{6}{22}, \text{PBAP} = \frac{16}{22}$$

$$H(T) = -\sum_j P_j \log_2 (P_j)$$

$$-\frac{6}{22}\log_2\left(\frac{6}{22}\right) - \frac{16}{22}\log_2\left(\frac{16}{22}\right)$$

$$= -0.273 \log_2(0.273) - 0.727 \log_2(0.727)$$

$$= -0.273\,(-1.873) - 0.727\,(0.459) = 0.845 \text{ bit}$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.845 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 5 records have 'Small family' size and 17 records have 'Large family' size.

$$P_{small} = \frac{5}{22}, P_{large} = \frac{17}{22}$$

Entropy for small family size

5 records (2 good, 3 bad)

$$-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$= -0.4 \log_2(0.4) - 0.6 \log_2(0.6)$$

$$= -0.4\,(-1.321) - 0.6\,(-0.736) = 0.97$$

Entropy for large family size

17 records (4 good, 13 bad)

$$-\frac{4}{17}\log_2\left(\frac{4}{17}\right) - \frac{13}{17}\log_2\left(\frac{13}{17}\right)$$

$$= -0.235 \log_2(0.235) - 0.765 \log_2(0.765)$$

$$= -0.235\,(-2.089) - 0.765\,(-0.386) = 0.786$$

Combine the entropies of these two subsets

$$\frac{5}{22}(0.97) + \frac{17}{22}(0.786)$$

$$= 0.227\,(0.97) + 0.773\,(0.786) = 0.828$$

The information gain represented by the split on the size of family attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.845 - 0.828 = **0.017 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), 12 records have 'Educated' status and 10 records have 'Uneducated' status.

$$P_{\text{educated}} = \frac{12}{22},\ P_{\text{uneducated}} = \frac{10}{22}$$

Entropy for educated status

12 records (4 good, 8 bad)

$$-\frac{4}{12}\log_2\left(\frac{4}{12}\right) - \frac{8}{12}\log_2\left(\frac{8}{12}\right)$$

$$= -0.333\log_2(0.333) - 0.667\log_2(0.667)$$

$$= -0.333(-1.586) - 0.667(-0.584) = 0.917$$

Entropy for uneducated status

10 records (2 good, 8 bad)

$$-\frac{2}{10}\log_2\left(\frac{2}{10}\right) - \frac{8}{10}\log_2\left(\frac{8}{10}\right)$$

$$= -0.2\log_2(0.2) - 0.8\log_2(0.8)$$

$$= -0.2(-2.321) - 0.8(-0.321) = 0.721$$

Combine the entropies of these two subsets,

$$\frac{12}{22}(0.917) + \frac{10}{22}(0.721)$$

$$= 0.455(0.721) + 0.545(0.917) = 0.827$$

The information gain represented by the split on the Parents education level attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.845 - 0.827 = **0.018 bit**

**Parent's employment status**

For candidate split 3 (Parents employment status), all 22 records have 'Unemployed' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 22 records have 'Single' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 15 records have low income and 7 records have high income.

$$P_{\text{low income}} = \frac{15}{22}, P_{\text{high income}} = \frac{7}{22}$$

Entropy for low income

15 records (4 good, 11 bad)

$$-\frac{4}{15}\log_2\left(\frac{4}{15}\right) - \frac{11}{15}\log_2\left(\frac{11}{15}\right)$$

$$= -0.267\log_2(0.267) - 0.733\log_2(0.733)$$

$$= -0.267(-1.905) - 0.733(-0.448) = 0.837$$

Entropy for high income

7 records (2 good, 5 bad)

$$-\frac{2}{7}\log_2\left(\frac{2}{7}\right) - \frac{5}{7}\log_2\left(\frac{5}{7}\right)$$

$$= -0.286\log_2(0.286) - 0.714\log_2(0.714)$$

$$= -0.286(-1.806) - 0.714(-0.486) = 0.863$$

Combine the entropies of these two subsets,

$$\frac{15}{22}(0.837) + \frac{7}{22}(0.863)$$

$$= 0.681(0.837) + 0.318(0.863) = 0.844$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.845 – 0.844 = **0.011 bit**

Table 4.10 summarizes the information gain for each candidate split at decision node 4. Candidate split 1, Parents education level has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.6

Table 4.10 Information Gain for each candidate split at decision node 4

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.017 bits |
| 2 | Parent's education level = Educated<br>Parent's education level = Uneducated | 0.018 bits |
| 3 | Parent's employment status = Unemployed | 0.000 bits |
| 4 | Parents marital status = Single | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.011 bits |



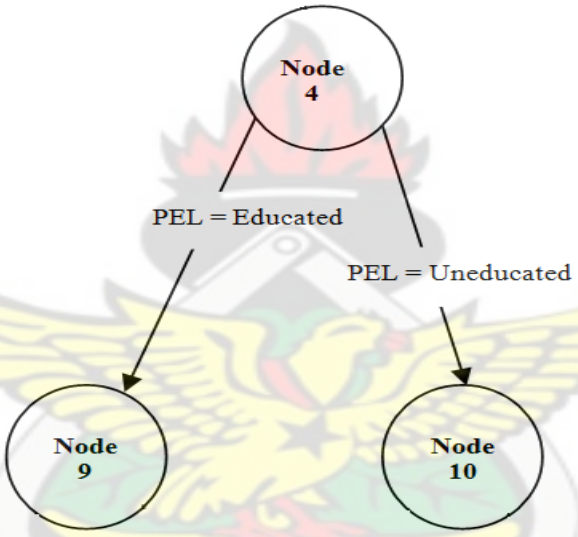


Figure 4.6 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of two new decision nodes. The 12 records at decision node 9 (Parents education level = educated) contain both good and bad academic performance. In the same vein, the 10 records at decision node 10 (Parents education level = uneducated) also contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.6 Candidate split at decision node 5

Determining the optimal split for decision node 5, containing records (1, 6, 9, 13, 15, 24, 26, 27, 29, 31, 33, 37, 43, 49, 51, 55, 58, 62, 64, 70, 81, 84, 96, 98) as indicated in Table 4.11 below.

Table 4.11 Training set of records available at decision node 5

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **1** | Small | Educated | Employed | Married | High | Good |
| **6** | Large | Educated | Employed | Married | Low | Good |
| **9** | Large | Educated | Employed | Married | Low | Good |
| **13** | Small | Educated | Unemployed | Married | High | Bad |
| **15** | Small | Educated | Unemployed | Married | High | Good |
| **24** | Small | Educated | Employed | Married | Low | Bad |
| **26** | Small | Educated | Employed | Married | High | Good |
| **27** | Large | Educated | Unemployed | Married | Low | Good |
| **29** | Large | Educated | Employed | Married | Low | Bad |
| **31** | Small | Educated | Employed | Married | High | Good |
| **33** | Small | Educated | Employed | Married | High | Bad |
| **37** | Small | Educated | Employed | Married | High | Bad |
| **43** | Small | Educated | Employed | Married | High | Good |
| **49** | Large | Educated | Employed | Married | Low | Good |
| **51** | Large | Educated | Employed | Married | Low | Good |
| **55** | Small | Educated | Unemployed | Married | Low | Bad |
| **58** | Large | Educated | Unemployed | Married | Low | Good |
| **62** | Large | Educated | Unemployed | Married | High | Bad |
| **64** | Large | Educated | Unemployed | Married | High | Bad |
| **70** | Large | Educated | Unemployed | Married | Low | Bad |
| **81** | Small | Educated | Unemployed | Married | High | Bad |
| **84** | Large | Educated | Employed | Married | Low | Bad |
| **96** | Small | Educated | Employed | Married | Low | Good |
| **98** | Small | Educated | Employed | Married | High | Good |

Entropy before splitting

13 of the 24 records are classified as good academic performance, with the remaining 11 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{13}{24}, \text{PBAP} = \frac{11}{24}$$

$$H(T) = -\sum_{j} P\,j \log_2 (P\,j)$$

$$-\frac{13}{24}\log_2\left(\frac{13}{24}\right) - \frac{11}{24}\log_2\left(\frac{11}{24}\right)$$

$$= -0.542 \log_2(0.542) - 0.458 \log_2(0.458))$$

$$= -0.542\,(-0.884) - 0.458\,(-1.126) = 0.994 \text{ bit}$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.994 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 13 records have 'Small family' size and 11 records have 'Large family' size.

$$P_{small} = \frac{13}{24}, P_{large} = \frac{11}{24}$$

Entropy for small family size

13 records (7 good, 6 bad)

$$-\frac{7}{13}\log_2\left(\frac{7}{13}\right)-\frac{6}{13}\log_2\left(\frac{6}{13}\right)$$

$$=-0.538\log_2(0.538)-0.462\log_2(0.462)$$

$$=-0.538(-0.894)-0.462(-1.114)=0.995$$

Entropy for large family size

11 records (6 good, 5 bad)

$$-\frac{6}{11}\log_2\left(\frac{6}{11}\right)-\frac{5}{11}\log_2\left(\frac{5}{11}\right)$$

$$=-0.545\log_2(0.545)-0.455\log_2(0.455)$$

$$=-0.545(-0.875)-0.455(-1.136)=0.994$$

Combine the entropies of these two subsets

$$\frac{13}{24}(0.995)+\frac{11}{24}(0.994)$$

$$=0.542(0.995)+0.458(0.994)=0.994$$

The information gain represented by the split on the size of family attribute is;

$$\text{gain}(S)=H(T)-H_S(T)$$

0.994 - 0.994 = **0.000 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), all 24 records have 'educated' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), 9 records have 'Unemployed' status and 15 records have 'Employed' status.

$$P_{\text{unemployed}}=\frac{9}{24},\ P_{\text{employed}}=\frac{15}{24}$$

Entropy for unemployed status

9 records (3 good, 6 bad)

$$-\frac{3}{9}\log_2\left(\frac{3}{9}\right)-\frac{6}{9}\log_2\left(\frac{6}{9}\right)$$

$$= -0.333 \log_2(0.333) - 0.667 \log_2(0.667)$$

$$= -0.333(-1.586) - 0.667(-0.584) = 0.917$$

Entropy for employed status

15 records (10 good, 5 bad)

$$-\frac{10}{15}\log_2\left(\frac{10}{15}\right) - \frac{5}{15}\log_2\left(\frac{5}{15}\right)$$

$$= -0.667 \log_2(0.667) - 0.333 \log_2(0.333)$$

$$= -0.667(-0.584) - 0.333(-1.586) = 0.917$$

Combine the entropies of these two subsets,

$$\frac{9}{24}(0.917) + \frac{15}{24}(0.917)$$

$$= 0.375(0.917) + 0.625(0.917) = 0.917$$

The information gain represented by the split on the Parents employment status attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.994 – 0.917 = **0.077 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 24 records have 'married' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 12 records have low income and 12 records have high income.

$$P_{\text{low income}} = \frac{12}{24}, P_{\text{high income}} = \frac{12}{24}$$

Entropy for low income

12 records (7 good, 5 bad)

$$-\frac{7}{12}\log_2\left(\frac{7}{12}\right) - \frac{5}{12}\log_2\left(\frac{5}{12}\right)$$

$$= -0.583 \log_2(0.583) - 0.417 \log_2(0.417)$$

$$= -0.583(-0.778) - 0.417(-1.261) = 0.979$$

Entropy for high income

12 records (6 good, 6 bad)

$$-\frac{6}{12}\log_2\left(\frac{6}{12}\right)-\frac{6}{12}\log_2\left(\frac{6}{12}\right)$$

$$=-0.5\log_2(0.5)-0.5\log_2(0.5)$$

$$=-0.5(-1)-0.5(-1)=1$$

Combine the entropies of these two subsets,

$$\frac{12}{24}(0.979)+\frac{12}{24}(1)$$

$$=0.5(0.979)+0.5(1)=0.989$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

$$0.994 - 0.989 = \mathbf{0.005\ bit}$$

Table 4.12 summarizes the information gain for each candidate split at decision node 5. Candidate split 3, Parents employment status has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.7

Table 4.12 Information Gain for each candidate split at decision node 5

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.000 bits |
| 2 | Parent's education level = Educated | 0.000 bits |
| 3 | Parent's employment status = Employed<br>Parent's employment status = Unemployed | 0.077 bits |
| 4 | Parents marital status = Married | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.005 bits |

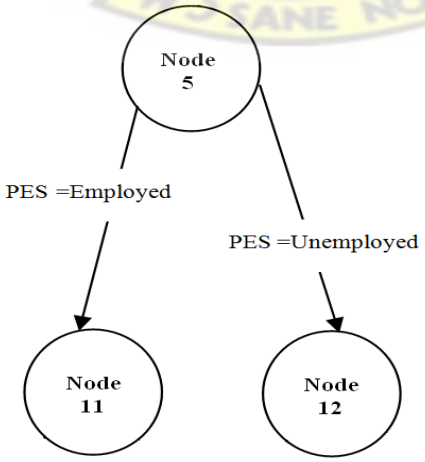


Figure 4.7 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of two new decision nodes. The 15 records at decision node 11 (Parents employment status = employed) contain both good and bad academic performance. In the same vein, the 9 records at decision node 12 (Parents employment status = unemployed) also contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.7 Candidate split at decision node 6

Determining the optimal split for decision node 6, containing records (3, 5, 7, 11, 16, 18, 20, 21, 35, 39, 41, 45, 47, 53, 56, 57, 59, 65, 67, 69, 72, 74, 76, 77, 79, 82, 85, 87, 89, 91, 92, 94, 100) as indicated in Table 4.13 below.

Table 4.13 Training set of records available at decision node 6

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **3** | Large | Uneducated | Employed | Married | High | Bad |
| **5** | Small | Uneducated | Employed | Married | High | Bad |
| **7** | Small | Uneducated | Employed | Married | Low | Good |
| **11** | Small | Uneducated | Unemployed | Married | Low | Good |
| **16** | Large | Uneducated | Employed | Married | Low | Good |
| **18** | Small | Uneducated | Employed | Married | Low | Bad |
| **20** | Small | Uneducated | Unemployed | Married | High | Good |
| **21** | Large | Uneducated | Employed | Married | Low | Good |
| **35** | Small | Uneducated | Unemployed | Married | High | Bad |
| **39** | Large | Uneducated | Employed | Married | High | Bad |
| **41** | Large | Uneducated | Unemployed | Married | Low | Good |
| **45** | Large | Uneducated | Unemployed | Married | Low | Good |
| **47** | Small | Uneducated | Unemployed | Married | High | Good |
| **53** | Large | Uneducated | Employed | Married | Low | Bad |
| **56** | Large | Uneducated | Employed | Married | High | Good |
| **57** | Small | Uneducated | Employed | Married | Low | Bad |
| **59** | Large | Uneducated | Employed | Married | High | Good |
| **65** | Large | Uneducated | Employed | Married | Low | Good |
| **67** | Small | Uneducated | Employed | Married | High | Good |
| **69** | Small | Uneducated | Employed | Married | High | Good |
| **72** | Large | Uneducated | Unemployed | Married | Low | Good |
| **74** | Small | Uneducated | Unemployed | Married | High | Good |
| **76** | Small | Uneducated | Employed | Married | Low | Bad |
| **77** | Small | Uneducated | Unemployed | Married | High | Good |
| **79** | Large | Uneducated | Employed | Married | High | Bad |
| **82** | Large | Uneducated | Employed | Married | Low | Good |
| **85** | Small | Uneducated | Unemployed | Married | High | Good |
| **87** | Large | Uneducated | Unemployed | Married | High | Good |
| **89** | Small | Uneducated | Unemployed | Married | Low | Bad |
| **91** | Small | Uneducated | Unemployed | Married | High | Bad |
| **92** | Small | Uneducated | Employed | Married | Low | Good |
| **94** | Small | Uneducated | Employed | Married | Low | Bad |
| **100** | Small | Uneducated | Employed | Married | Low | Bad |

Entropy before splitting

20 of the 33 records are classified as good academic performance, with the remaining 13 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{20}{33}, \text{PBAP} = \frac{13}{33}$$

$$H(T) = -\sum_{j} P_j \log_2 (P_j)$$

$$-\frac{20}{33}\log_2\left(\frac{20}{33}\right) - \frac{13}{33}\log_2\left(\frac{13}{33}\right)$$

$$= -0.606 \log_2(0.606) - 0.394 \log_2(0.394)$$

$$= -0.606\,(-0.722) - 0.394\,(-1.343) = 0.967$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.967 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 19 records have 'Small family' size and 14 records have 'Large family' size.

$$P_{small} = \frac{19}{33},\ P_{large} = \frac{14}{33}$$

Entropy for small family size

19 records (10 good, 9 bad)

$$-\frac{10}{19}\log_2\left(\frac{10}{19}\right) - \frac{9}{19}\log_2\left(\frac{9}{19}\right)$$

$$= -0.526 \log_2(0.526) - 0.474 \log_2(0.474)$$

$$= -0.526\,(-0.927) - 0.474\,(-1.077) = 0.998$$

Entropy for large family size

14 records (10 good, 4 bad)

$$-\frac{10}{14}\log_2\left(\frac{10}{14}\right) - \frac{4}{14}\log_2\left(\frac{4}{14}\right)$$

$$= -0.714 \log_2(0.714) - 0.286 \log_2(0.286)$$

$$= -0.714\,(-0.486) - 0.286\,(-1.806) = 0.863$$

Combine the entropies of these two subsets

$$\frac{19}{33}(0.998) + \frac{14}{33}(0.863)$$

$$= 0.576\,(0.998) + 0.424\,(0.863) = 0.940$$

The information gain represented by the split on the size of family attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.967 - 0.940 = **0.027 bit**

**Parent's education level**

For candidate split 2 (Parents education level), all 33 records have 'uneducated' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), 13 records have 'Unemployed' status and 20 records have 'Employed' status.

$$P_{unemployed} = \frac{13}{33}, P_{employed} = \frac{20}{33}$$

Entropy for unemployed status

13 records (10 good, 3 bad)

$$-\frac{10}{13}\log_2\left(\frac{10}{13}\right) - \frac{3}{13}\log_2\left(\frac{3}{13}\right)$$

$$= -0.769\log_2(0.769) - 0.231\log_2(0.231)$$

$$= -0.769(-0.378) - 0.231(-2.114) = 0.779$$

Entropy for employed status

20 records (10 good, 10 bad)

$$-\frac{10}{20}\log_2\left(\frac{10}{20}\right) - \frac{10}{20}\log_2\left(\frac{10}{20}\right)$$

$$= -0.5\log_2(0.5) - 0.5\log_2(0.5)$$

$$= -0.5(-1) - 0.5(-1) = 1$$

Combine the entropies of these two subsets,

$$\frac{13}{33}(0.779) + \frac{20}{33}(1)$$

$$= 0.394(0.779) + 0.606(1.00) = 0.912$$

The information gain represented by the split on the Parents employment status attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.967 – 0.912 = **0.055 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 33 records have 'married' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same

number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 17 records have low income and 16 records have high income.

$$P_{\text{low income}} = \frac{17}{33}, P_{\text{high income}} = \frac{16}{33}$$

Entropy for low income

17 records (11 good, 6 bad)

$$-\frac{11}{17}\log_2\left(\frac{11}{17}\right) - \frac{6}{17}\log_2\left(\frac{6}{17}\right)$$

$$= -0.647\log_2(0.647) - 0.353\log_2(0.353)$$

$$= -0.647(-0.628) - 0.353(-1.502) = 0.936$$

Entropy for high income

16 records (10 good, 6 bad)

$$-\frac{10}{16}\log_2\left(\frac{10}{16}\right) - \frac{6}{16}\log_2\left(\frac{6}{16}\right)$$

$$= -0.625\log_2(0.625) - 0.375\log_2(0.375)$$

$$= -0.625(-0.678) - 0.375(-1.415) = 0.954$$

Combine the entropies of these two subsets,

$$\frac{17}{33}(0.936) + \frac{16}{33}(0.954)$$

$$= 0.515(0.936) + 0.485(0.954) = 0.944$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.967 – 0.944 = **0.023 bit**

Table 4.14 summarizes the information gain for each candidate split at decision node 6. Candidate split 3, Parents employment status has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.8

Table 4.14 Information Gain for each candidate split at decision node 6

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.027 bits |
| 2 | Parent's education level = Uneducated | 0.000 bits |
| 3 | Parent's employment status = Employed<br>Parent's employment status = Unemployed | 0.055 bits |
| 4 | Parents marital status = Married | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.023 bits |



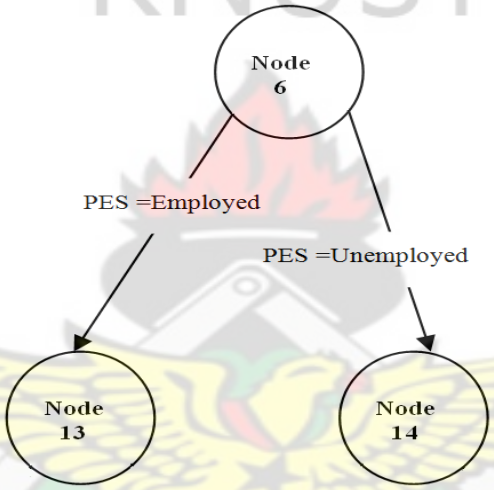


Figure 4.8 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of two new decision nodes. The 20 records at decision node 13 (Parents employment status = employed) contain both good and bad academic performance. In the same vein, the 13 records at decision node 14 (Parents employment status = unemployed) also contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.8 Candidate split at decision node 7

Determining the optimal split for decision node 7, containing records (23, 28, 46, 60, 63, 86, 88) as indicate d in Table 4.15 below.

Table 4.15 Training set of records available at decision node 7

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **23** | Small | Uneducated | Employed | Single | High | Good |
| **28** | Small | Uneducated | Employed | Single | High | Bad |
| **46** | Small | Educated | Employed | Single | High | Bad |
| **60** | Small | Educated | Employed | Single | High | Bad |
| **63** | Small | Uneducated | Employed | Single | Low | Good |
| **86** | Small | Educated | Employed | Single | Low | Bad |
| **88** | Small | Educated | Employed | Single | Low | Bad |

Entropy before splitting

2 of the 7 records are classified as good academic performance, with the remaining 5 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{2}{7}, \text{PBAP} = \frac{5}{7}$$

$$H(T) = -\sum_{j} P\,j \log_2 (P\,j)$$

$$-\frac{2}{7}\log_2\left(\frac{2}{7}\right) - \frac{5}{7}\log_2\left(\frac{5}{7}\right)$$

$$= -0.285 \log_2(0.285) - 0.714 \log_2(0.714)$$

$$= -0.285\,(-1.810) - 0.714\,(-0.486) = 0.863$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.863 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (Size of family), all 7 records have 'small' family status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), 4 records have 'Educated' status and 3 records have 'Uneducated' status.

$$P_{educated} = \frac{4}{7}, P_{uneducated} = \frac{3}{7}$$

Entropy for educated status

4 records (0 good, 4 bad)

$$-\frac{0}{4}\log_2\left(\frac{0}{4}\right) - \frac{4}{4}\log_2\left(\frac{4}{4}\right)$$

$$= -0\log_2(0) - 1\log_2(0)$$

$$= -0\,(0) - 1\,(0) = 0$$

Entropy for uneducated status

3 records (2 good, 1 bad)

$$-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)$$

$$= -0.667 \log_2(0.667) - 0.333 \log_2(0.333)$$

$$= -0.667\,(-0.584) - 0.333\,(-1.586) = 0.917$$

Combine the entropies of these two subsets,

$$\frac{4}{7}(0) + \frac{3}{7}(0.917)$$

$$= 0.571\,(0) + 0.429\,(0.917) = 0.393$$

The information gain represented by the split on the Parents employment status attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.863 – 0.393 = **0.47 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), all 7 records have 'employed' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 7 records have 'single' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 3 records have low income and 4 records have high income.

$$P_{\text{low income}} = \frac{3}{7},\ P_{\text{high income}} = \frac{4}{7}$$

Entropy for low income

3 records (1 good, 2 bad)

$$-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$$

$$= -0.333\log_2(0.333) - 0.667\log_2(0.667)$$

$$= -0.333\,(-1.586) - 0.667\,(-0.584) = 0.917$$

Entropy for high income

4 records (1 good, 3 bad)

$$-\frac{1}{4}\log_2\left(\frac{1}{4}\right)-\frac{3}{4}\log_2\left(\frac{3}{4}\right)$$

$$=-0.25\log_2(0.25)-0.75\log_2(0.75)$$

$$=-0.25(-2)-0.75(-0.415)=0.811$$

Combine the entropies of these two subsets,

$$\frac{3}{7}(0.917)+\frac{4}{7}(0.811)$$

$$=0.428(0.917)+0.571(0.811)=0.856$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

$$0.863 - 0.856 = \mathbf{0.007\ bit}$$

Table 4.16 summarizes the information gain for each candidate split at decision node 7. Candidate split 2, Parents education level status has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.9

Table 4.16 Information Gain for each candidate split at decision node 7

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small | 0.000 bits |
| 2 | Parent's education level = Educated<br>Parent's education level = Uneducated | 0.047 bits |
| 3 | Parent's employment status = Employed<br>Parent's employment status = Unemployed | 0.000 bits |
| 4 | Parents marital status = Single | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.007 bits |

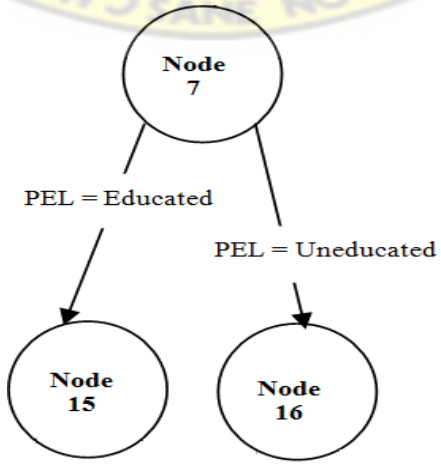


Figure 4.9 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of two new decision nodes. The 3 records at decision node 15 (Parents education level = educated) contain both good and bad academic performance. In the same vein, the 4 records at decision node 16 (Parents education level = uneducated) also contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.9 Candidate split at decision node 8

Determining the optimal split for decision node 8, containing records (10, 12, 14, 19, 34, 40, 44, 54, 61, 71, 73, 75, 80, 90) as indicated in Table 4.17 below

Table 4.17 Training set of records available at decision node 8

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **10** | Large | Educated | Employed | Single | High | Bad |
| **12** | Large | Educated | Employed | Single | High | Bad |
| **14** | Large | Uneducated | Employed | Single | Low | Bad |
| **19** | Large | Educated | Employed | Single | High | Bad |
| **34** | Large | Educated | Employed | Single | Low | Good |
| **40** | Large | Educated | Employed | Single | High | Good |
| **44** | Large | Educated | Employed | Single | Low | Bad |
| **54** | Large | Educated | Employed | Single | High | Good |
| **61** | Large | Uneducated | Employed | Single | Low | Bad |
| **71** | Large | Educated | Employed | Single | High | Bad |
| **73** | Large | Educated | Employed | Single | Low | Bad |
| **75** | Large | Educated | Employed | Single | Low | Bad |
| **80** | Large | Educated | Employed | Single | Low | Good |
| **90** | Large | Educated | Employed | Single | Low | Good |

Entropy before splitting

5 of the 14 records are classified as good academic performance, with the remaining 9 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{5}{14}, \text{PBAP} = \frac{9}{14}$$

$$H(T) = -\sum_{j} P\,j \log_2 (P\,j)$$

$$-\frac{5}{14}\log_2\left(\frac{5}{14}\right) - \frac{9}{14}\log_2\left(\frac{9}{14}\right)$$

$$= -0.357 \log_2(0.357) - 0.643 \log_2(0.643)$$

$$= -0.357(-1.486) - 0.643(-0.637) = 0.940$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.94 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (Size of family), all 14 records have 'large' family status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), 12 records have 'educated' status and 2 records have 'uneducated' status.

$$P_{educated} = \frac{12}{14}, P_{uneducated} = \frac{2}{14}$$

Entropy for educated status

12 records (5 good, 7 bad)

$$-\frac{5}{12}\log_2\left(\frac{5}{12}\right) - \frac{7}{12}\log_2\left(\frac{7}{12}\right)$$

$$= -0.417\log_2(0.417) - 0.583\log_2(0.583)$$

$$= -0.417\,(-1.261) - 0.583\,(-0.778) = 0.979$$

Entropy for uneducated status

2 records (0 good, 2 bad)

$$-\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)$$

$$= -0\log_2(0) - 1\log_2(1)$$

$$= -0\,(0) - 1\,(0) = 0$$

Combine the entropies of these two subsets,

$$\frac{12}{14}(0.979) + \frac{2}{14}(0)$$

$$= 0.857(0.979) + 0.143\,(0) = 0.839$$

The information gain represented by the split on the Parents education level attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.94 - 0.839 = **0.101 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), all 14 records have 'employed' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use

the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 14 records have 'single' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 8 records have low income and 6 records have high income.

$$P_{\text{low income}} = \frac{8}{14}, P_{\text{high income}} = \frac{6}{14}$$

Entropy for low income

8 records (3 good, 5 bad)

$$-\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right)$$

$$= -0.375\log_2(0.375) - 0.625\log_2(0.625)$$

$$= -0.375(-1.415) - 0.625(-0.678) = 0.954$$

Entropy for high income

6 records (2 good, 4 bad)

$$-\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\log_2\left(\frac{4}{6}\right)$$

$$= -0.333\log_2(0.333) - 0.667\log_2(0.667)$$

$$= -0.333(-1.586) - 0.667(-0.584) = 0.917$$

Combine the entropies of these two subsets,

$$\frac{8}{14}(0.954) + \frac{6}{14}(0.917)$$

$$= 0.571(0.954) + 0.429(0.917) = 0.938$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.94 – 0.938 = **0.002 bit**

Table 4.18 summarizes the information gain for each candidate split at decision node 8. Candidate split 2, Parents education level has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.10

Table 4.18 Information Gain for each candidate split at decision node 8

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Large | 0.000 bits |
| 2 | Parent's education level = Educated<br>Parent's education level = Uneducated | 0.101 bits |
| 3 | Parent's employment status = Employed | 0.000 bits |
| 4 | Parents marital status = Single | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.002 bits |



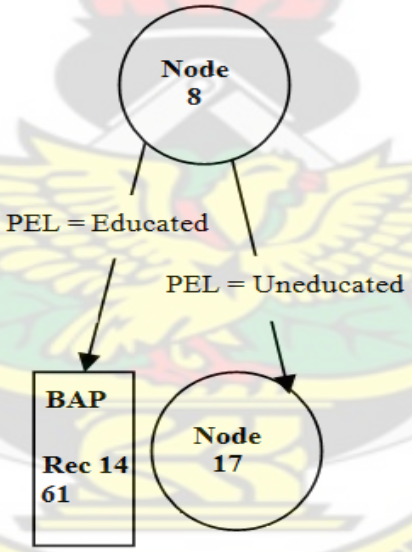


Figure 4.10 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of one terminal leaf node and one new decision node. Records with (Parents education level = educated) have only bad academic performance therefore no further splits are required. The 12 records at decision node 17 (Parents education level = uneducated) contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.10 Candidate split at decision node 9

Determining the optimal split for decision node 9, containing records (4, 8, 17, 22, 36, 52, 66, 68, 78, 93, 95, 99) as indicated in Table 4.19 below.

Table 4.19 Training set of records available at decision node 9

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **4** | Large | Educated | Unemployed | Single | Low | Good |
| **8** | Large | Educated | Unemployed | Single | High | Bad |
| **17** | Large | Educated | Unemployed | Single | High | Bad |
| **22** | Large | Educated | Unemployed | Single | Low | Bad |
| **36** | Large | Educated | Unemployed | Single | Low | Good |
| **52** | Small | Educated | Unemployed | Single | High | Good |
| **66** | Small | Educated | Unemployed | Single | Low | Bad |
| **68** | Large | Educated | Unemployed | Single | Low | Bad |
| **78** | Large | Educated | Unemployed | Single | Low | Bad |
| **93** | Large | Educated | Unemployed | Single | Low | Good |
| **95** | Large | Educated | Unemployed | Single | High | Bad |
| **99** | Large | Educated | Unemployed | Single | Low | Bad |

Entropy before splitting

4 of the 12 records are classified as good academic performance, with the remaining 8 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{4}{12}, \text{PBAP} = \frac{8}{12}$$

$$H(T) = -\sum_{j} P\,j \log_2 (P\,j)$$

$$-\frac{4}{12}\log_2\left(\frac{4}{12}\right) - \frac{8}{12}\log_2\left(\frac{8}{12}\right)$$

$$= -0.333 \log_2(0.333) - 0.667 \log_2(0.667)$$

$$= -0.333(-1.586) - 0.667(-0.584) = 0.917$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.917 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 2 records have 'Small family' size and 10 records have 'Large family' size.

$$P_{small} = \frac{2}{12}, P_{large} = \frac{10}{12}$$

Entropy for small family size

2 records (1 good, 1 bad)

$$-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$= -0.5 \log_2(0.5) - 0.5 \log_2(0.5)$$

$$= -0.5(-1) - 0.5(-1) = 1$$

Entropy for large family size

10 records (3 good, 7 bad)

$$-\frac{3}{10}\log_2\left(\frac{3}{10}\right)-\frac{7}{10}\log_2\left(\frac{7}{10}\right)$$

$$=-0.3\log_2(0.3)-0.7\log_2(0.7)$$

$$=-0.3(-1.736)-0.7(-0.514)=0.881$$

Combine the entropies of these two subsets

$$\frac{2}{12}(1)+\frac{10}{12}(0.881)$$

$$=0.167(1)+0.833(0.881)=0.901$$

The information gain represented by the split on the size of family attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.917 - 0.901 = **0.016 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), all 12 records have 'educated' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), all 12 records have 'unemployed' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 12 records have 'single' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 8 records have low income and 4 records have high income.

$$P_{\text{low income}} = \frac{8}{12},\ P_{\text{high income}} = \frac{4}{12}$$

Entropy for low income

8 records (3 good, 5 bad)

$$-\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right)$$

$$= -0.375\log_2(0.375) - 0.625\log_2(0.625)$$

$$= -0.375(-1.415) - 0.625(-0.678) = 0.954$$

Entropy for high income

4 records (1 good, 3 bad)

$$-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)$$

$$= -0.25\log_2(0.25) - 0.75\log_2(0.75)$$

$$= -0.25(-2) - 0.75(-0.415) = 0.811$$

Combine the entropies of these two subsets,

$$\frac{8}{12}(0.954) + \frac{4}{12}(0.811)$$

$$= 0.667(0.954) + 0.333(0.811) = 0.906$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

$$0.917 - 0.906 = \mathbf{0.011\ bit}$$

Table 4.20 summarizes the information gain for each candidate split at decision node 9. Candidate split 1, Size of family has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.11

Table 4.20 Information Gain for each candidate split at decision node 9

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.016 bits |
| 2 | Parent's education level = Educated | 0.000 bits |
| 3 | Parent's employment status = Unemployed | 0.000 bits |
| 4 | Parents marital status = Single | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.011 bits |

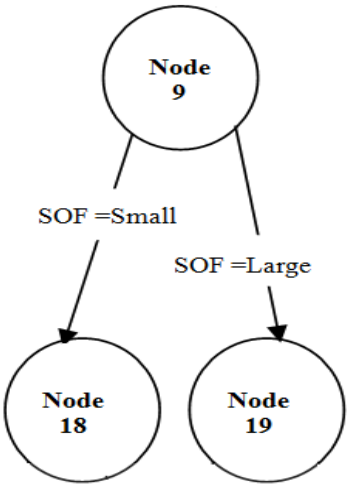


Figure 4.11 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of two new decision nodes. The 10 records at decision node 18 (Size of family = small) contain both good and bad academic performance. In the same vein, the 2 records at decision node 19 (Size of family = large) also contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.11 Candidate split at decision node 10

Determining the optimal split for decision node 10, containing records (2, 25, 30, 32, 38, 42, 48, 50, 83, 97) as indicated in Table 4.21 below.

Table 4.21 Training set of records available at decision node 10

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **2** | Large | Uneducated | Unemployed | Single | Low | Bad |
| **25** | Large | Uneducated | Unemployed | Single | High | Good |
| **30** | Large | Uneducated | Unemployed | Single | Low | Bad |
| **32** | Large | Uneducated | Unemployed | Single | Low | Bad |
| **38** | Small | Uneducated | Unemployed | Single | Low | Good |
| **42** | Large | Uneducated | Unemployed | Single | Low | Bad |
| **48** | Large | Uneducated | Unemployed | Single | Low | Bad |
| **50** | Small | Uneducated | Unemployed | Single | High | Bad |
| **83** | Small | Uneducated | Unemployed | Single | High | Bad |
| **97** | Large | Uneducated | Unemployed | Single | Low | Bad |

Entropy before splitting

2 of the 10 records are classified as good academic performance, with the remaining 8 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{2}{10}, \text{PBAP} = \frac{8}{10}$$

$$H(T) = -\sum_{j} P\,j \log_2 (P\,j)$$

$$-\frac{2}{10}\log_2\left(\frac{2}{10}\right)-\frac{8}{10}\log_2\left(\frac{8}{10}\right)$$

$$=-0.2\log_2(0.2)-0.8\log_2(0.8)$$

$$=-0.2(-2.321)-0.8(-0.321)=0.721 \text{ bit}$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.721 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 3 records have 'Small family' size and 7 records have 'Large family' size.



$$P_{small}=\frac{3}{10},\ P_{large}=\frac{7}{10}$$

Entropy for small family size

3 records (1 good, 2 bad)

$$-\frac{1}{3}\log_2\left(\frac{1}{3}\right)-\frac{2}{3}\log_2\left(\frac{2}{3}\right)$$

$$=-0.333\log_2(0.333)-0.667\log_2(0.667)$$

$$=-0.333(-1.586)-0.667(-0.584)=0.917$$

Entropy for large family size

7 records (1 good, 6 bad)

$$-\frac{1}{7}\log_2\left(\frac{1}{7}\right)-\frac{6}{7}\log_2\left(\frac{6}{7}\right)$$

$$=-0.143\log_2(0.143)-0.857\log_2(0.857)$$

$$=-0.143(-2.806)-0.857(-0.223)=0.592$$

Combine the entropies of these two subsets

$$\frac{3}{10}(0.917)+\frac{7}{10}(0.592)$$

$$=0.3(0.917)+0.7(0.592)=0.689$$

The information gain represented by the split on the size of family attribute is;

$$\text{gain}(S)=H(T)-H_S(T)$$

0.721 - 0.689 = **0.032 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), all 10 records have 'uneducated' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use

the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), all 10 records have 'unemployed' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents marital status**

For candidate split 4 (Parent's marital status), all 10 records have 'single' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 7 records have low income and 3 records have high income.

$$P_{\text{low income}} = \frac{7}{10}, P_{\text{high income}} = \frac{3}{10}$$

Entropy for low income

7 records (1 good, 6 bad)

$$-\frac{1}{7}\log_2\left(\frac{1}{7}\right) - \frac{6}{7}\log_2\left(\frac{6}{7}\right)$$

$$= -0.143\log_2(0.143) - 0.857\log_2(0.857)$$

$$= -0.143(-2.806) - 0.857(-0.223) = 0.592$$

Entropy for high income

3 records (1 good, 2 bad)

$$-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$$

$$= -0.333\log_2(0.333) - 0.667\log_2(0.667)$$

$$= -0.333(-1.586) - 0.667(-0.584) = 0.917$$

Combine the entropies of these two subsets,

$$\frac{7}{10}(0.592) + \frac{3}{10}(0.917)$$

$$= 0.7(0.592) + 0.3(0.917) = 0.689$$

The information gain represented by the split on the Parents monthly income attribute is;

$$gain(S) = H(T) - H_S(T)$$

$$0.721 - 0.689 = \mathbf{0.032\ bit}$$

Table 4.22 summarizes the information gain for each candidate split at decision node 10. Candidate split 1 and 5 have the same number of bits, but the first one encountered (size of family) is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.12

Table 4.22 Information Gain for each candidate split at decision node 10

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.032 bits |
| 2 | Parent's education level = Uneducated | 0.000 bits |
| 3 | Parent's employment status = Employed | 0.000 bits |
| 4 | Parents marital status = Single | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.032 bits |

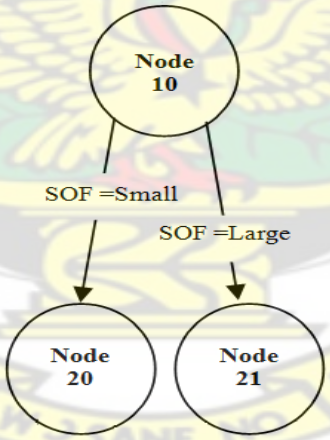


Figure 4.12 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of two new decision nodes. The 7 records at decision node 20 (Size of family = small) contain both good and bad academic performance. In the same vein, the 3 records at decision node 21 (Size of family = large) also contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.12 Candidate split at decision node 11

Determining the optimal split for decision node 11, containing records (1, 6, 9, 24, 26, 29, 31, 33, 37, 43, 49, 51, 84, 96, 98) as indicated in Table 4.23 below.

Table 4.23 Training set of records available at decision node 11

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **1** | Small | Educated | Employed | Married | High | Good |
| **6** | Large | Educated | Employed | Married | Low | Good |
| **9** | Large | Educated | Employed | Married | Low | Good |
| **24** | Small | Educated | Employed | Married | Low | Bad |
| **26** | Small | Educated | Employed | Married | High | Good |
| **29** | Large | Educated | Employed | Married | Low | Bad |
| **31** | Small | Educated | Employed | Married | High | Good |
| **33** | Small | Educated | Employed | Married | High | Bad |
| **37** | Small | Educated | Employed | Married | High | Bad |
| **43** | Small | Educated | Employed | Married | High | Good |
| **49** | Large | Educated | Employed | Married | Low | Good |
| **51** | Large | Educated | Employed | Married | Low | Good |
| **84** | Large | Educated | Employed | Married | Low | Bad |
| **96** | Small | Educated | Employed | Married | Low | Good |
| **98** | Small | Educated | Employed | Married | High | Good |



Entropy before splitting

10 of the 15 records are classified as good academic performance, with the remaining 5 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{10}{15}, \text{PBAP} = \frac{5}{15}$$

$$H(T) = -\sum_{j} P\,j \log_2 (P\,j)$$

$$-\frac{10}{15}\log_2\left(\frac{10}{15}\right) - \frac{5}{15}\log_2\left(\frac{5}{15}\right)$$

$$= -0.667 \log_2(0.667) - 0.333 \log_2(0.333)$$

$$= -0.667(-0.584) - 0.333(-1.586) = 0.917$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.917 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 9 records have 'Small family' size and 6 records have 'Large family' size.

$$P_{small} = \frac{9}{15}, P_{large} = \frac{6}{15}$$

Entropy for small family size

9 records (6 good, 3 bad)

$$-\frac{6}{9}\log_2\left(\frac{6}{9}\right)-\frac{3}{9}\log_2\left(\frac{3}{9}\right)$$

$$=-0.667\log_2(0.667)-0.333\log_2(0.333)$$

$$=-0.667(-0.584)-0.333(-1.586)=0.917$$

Entropy for large family size

6 records (4 good, 2 bad)

$$-\frac{4}{6}\log_2\left(\frac{4}{6}\right)-\frac{2}{6}\log_2\left(\frac{2}{6}\right)$$

$$=-0.667\log_2(0.667)-0.333\log_2(0.333)$$

$$=-0.667(-0.584)-0.333(-1.586)=0.917$$

Combine the entropies of these two subsets

$$\frac{9}{15}(0.917)+\frac{6}{15}(0.917)$$

$$=0.6\,(0.917)+0.4\,(0.917)=0.917$$

The information gain represented by the split on the size of family attribute is;

$$\text{gain}(S)=H(T)-H_S(T)$$

0.917 - 0.917 = **0.000 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), all 15 records have 'educated' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), all 15 records have 'employed' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 15 records have 'married' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 8 records have low income and 7 records have high income.

$$P_{\text{low income}} = \frac{8}{15},\ P_{\text{high income}} = \frac{7}{15}$$

Entropy for low income

8 records (5 good, 3 bad)

$$-\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right)$$

$$= -0.625\log_2(0.625) - 0.375\log_2(0.375)$$

$$= -0.625(-0.678) - 0.375(-1.415) = 0.954$$

Entropy for high income

7 records (5 good, 2 bad)

$$-\frac{5}{7}\log_2\left(\frac{5}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right)$$

$$= -0.714\log_2(0.714) - 0.286\log_2(0.286)$$

$$= -0.714(-0.486) - 0.286(-1.806) = 0.863$$

Combine the entropies of these two subsets,

$$\frac{8}{15}(0.954) + \frac{7}{15}(0.862)$$

$$= 0.533(0.954) + 0.467(0.862) = 0.911$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

$0.917 - 0.911 =$ **0.006 bit**

Table 4.24 summarizes the information gain for each candidate split at decision node 11. Candidate split 5, Monthly income has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.13

Table 4.24 Information Gain for each candidate split at decision node 11

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.000 bits |
| 2 | Parent's education level = Educated | 0.000 bits |
| 3 | Parent's employment status = Employed | 0.000 bits |
| 4 | Parents marital status = Married | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.006 bits |



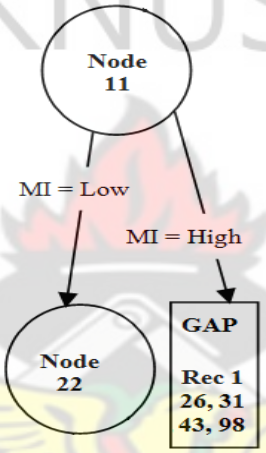


Figure 4.13 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of one terminal leaf node and one new decision node. Records with (Monthly income = high) have only good academic performance therefore no further splits are required. The 8 records at decision node 22 (Monthly income = low) contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.13 Candidate split at decision node 12

Determining the optimal split for decision node 12, containing records (13, 15, 27, 55, 58, 62, 64, 70, 81) as indicated in Table 4.25 below.

Table 4.25 Training set of records available at decision node 12

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **13** | Small | Educated | Unemployed | Married | High | Bad |
| **15** | Small | Educated | Unemployed | Married | High | Good |
| **27** | Large | Educated | Unemployed | Married | Low | Good |
| **55** | Small | Educated | Unemployed | Married | Low | Bad |
| **58** | Large | Educated | Unemployed | Married | Low | Good |
| **62** | Large | Educated | Unemployed | Married | High | Bad |
| **64** | Large | Educated | Unemployed | Married | High | Bad |
| **70** | Large | Educated | Unemployed | Married | Low | Bad |
| **81** | Small | Educated | Unemployed | Married | High | Bad |

Entropy before splitting

3 of the 9 records are classified as good academic performance, with the remaining 6 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{3}{9}, \text{PBAP} = \frac{6}{9}$$

$$H(T) = -\sum_j P\,j \log_2 (P\,j)$$

$$-\frac{3}{9}\log_2\left(\frac{3}{9}\right) - \frac{6}{9}\log_2\left(\frac{6}{9}\right)$$

$$= -0.333 \log_2(0.333) - 0.667 \log_2(0.667)$$

$$= -0.333(-1.586) - 0.667(-0.584) = 0.917$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.917 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 4 records have 'Small family' size and 5 records have 'Large family' size.

$$P_{small} = \frac{4}{9}, P_{large} = \frac{5}{9}$$

Entropy for small family size

4 records (1 good, 3 bad)

$$-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)$$

$$= -0.25 \log_2(0.25) - 0.75 \log_2(0.75)$$

$$= -0.25(-2) - 0.75(-0.415) = 0.811$$

Entropy for large family size

5 records (2 good, 3 bad)

$$-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$= -0.4 \log_2(0.4) - 0.6 \log_2(0.6)$$

$$= -0.4(-1.321) - 0.6(-0.736) = 0.97$$

Combine the entropies of these two subsets

$$\frac{4}{9}(0.811) + \frac{5}{9}(0.97)$$

$$= 0.444\,(0.811) + 0.556\,(0.97) = 0.899$$

The information gain represented by the split on the size of family attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.917 - 0.899 = **0.018 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), all 9 records have 'educated' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**



**Parent's employment status**

For candidate split 3 (Parent's employment status), all 9 records have 'unemployed' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 9 records have 'married' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 4 records have low income and 5 records have high income.

$$P_{\text{low income}} = \frac{4}{9},\ P_{\text{high income}} = \frac{5}{9}$$

Entropy for low income

4 records (2 good, 2 bad)

$$-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$$

$$= -0.5\log_2(0.5) - 0.5\log_2(0.5)$$

$$= -0.5\,(-1) - 0.5\,(-1) = 1$$

Entropy for high income

5 records (1 good, 4 bad)

$$-\frac{1}{5}\log_2\left(\frac{1}{5}\right)-\frac{4}{5}\log_2\left(\frac{4}{5}\right)$$

$$=-0.2\log_2(0.8)-0.8\log_2(0.2)$$

$$=-0.2(-2.321)-0.8(-0.321)=0.721$$

Combine the entropies of these two subsets,

$$\frac{4}{9}(1)+\frac{5}{9}(0.721)$$

$$=0.444(1)+0.556(0.721)=0.844$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

$$0.917 - 0.844 = \mathbf{0.073\ bit}$$

Table 4.26 summarizes the information gain for each candidate split at decision node 12. Candidate split 5, Monthly income has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.14

Table 4.26 Information Gain for each candidate split at decision node 12

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.013 bits |
| 2 | Parent's education level = Educated | 0.000 bits |
| 3 | Parent's employment status = Unemployed | 0.000 bits |
| 4 | Parents marital status = Married | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.073 bits |

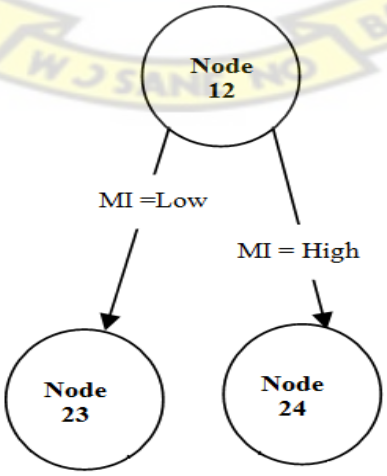


Figure 4.14 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of two new decision nodes. The 5 records at decision node 23 (Monthly income = low) contain both good and bad academic performance. In the same vein, the 4 records at decision node 24 (Monthly income = high) also contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.14 Candidate split at decision node 13

Determining the optimal split for decision node 13, containing records (3, 5, 7, 16, 18, 21, 39, 53, 56, 57, 59, 65, 67, 69, 76, 79, 82, 92, 94, 100) as indicated in Table 4.27 below.

Table 4.27 Training set of records available at decision node 13

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **3** | Large | Uneducated | Employed | Married | High | Bad |
| **5** | Small | Uneducated | Employed | Married | High | Bad |
| **7** | Small | Uneducated | Employed | Married | Low | Good |
| **16** | Large | Uneducated | Employed | Married | Low | Good |
| **18** | Small | Uneducated | Employed | Married | Low | Bad |
| **21** | Large | Uneducated | Employed | Married | Low | Good |
| **39** | Large | Uneducated | Employed | Married | High | Bad |
| **53** | Large | Uneducated | Employed | Married | Low | Bad |
| **56** | Large | Uneducated | Employed | Married | High | Good |
| **57** | Small | Uneducated | Employed | Married | Low | Bad |
| **59** | Large | Uneducated | Employed | Married | High | Good |
| **65** | Large | Uneducated | Employed | Married | Low | Good |
| **67** | Small | Uneducated | Employed | Married | High | Good |
| **69** | Small | Uneducated | Employed | Married | High | Good |
| **76** | Small | Uneducated | Employed | Married | Low | Bad |
| **79** | Large | Uneducated | Employed | Married | High | Bad |
| **82** | Large | Uneducated | Employed | Married | Low | Good |
| **92** | Small | Uneducated | Employed | Married | Low | Good |
| **94** | Small | Uneducated | Employed | Married | Low | Bad |
| **100** | Small | Uneducated | Employed | Married | Low | Bad |

Entropy before splitting

10 of the 20 records are classified as good academic performance, with the remaining 10 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{10}{20}, \text{PBAP} = \frac{10}{20}$$

$$H(T) = -\sum_{j} P\,j \log_2 (P\,j)$$

$$-\frac{10}{20}\log_2\left(\frac{10}{20}\right) - \frac{10}{20}\log_2\left(\frac{10}{20}\right)$$

$$= -0.5\log_2(0.5) - 0.5\log_2(0.5)$$

$$= -0.5(-1) - 0.5(-1) = 1$$

After the analysis, compare the entropy of each candidate split against H(T) = 1 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 10 records have 'Small family' size and 10 records have 'Large family' size.

$$P_{small} = \frac{10}{20}, P_{large} = \frac{10}{20}$$

Entropy for small family size

10 records (4 good, 6 bad)

$$-\frac{4}{10}\log_2\left(\frac{4}{10}\right) - \frac{6}{10}\log_2\left(\frac{6}{10}\right)$$

$$= -0.4\log_2(0.4) - 0.6\log_2(0.6)$$

$$= -0.4(-1.321) - 0.6(-0.736) = 0.97$$

Entropy for large family size

10 records (6 good, 4 bad)

$$-\frac{6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right)$$

$$= -0.6\log_2(0.6) - 0.4\log_2(0.4)$$

$$= -0.6(-0.736) - 0.4(-1.321) = 0.97$$

Combine the entropies of these two subsets

$$\frac{10}{20}(0.97) + \frac{10}{20}(0.97)$$

$$= 0.5(0.97) + 0.5(0.97) = 0.97$$

The information gain represented by the split on the size of family attribute is;

$$gain(S) = H(T) - H_S(T)$$

1 - 0.97 = **0.003 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), all 20 records have 'uneducated' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), all 20 records have 'employed' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 20 records have 'married' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 12 records have low income and 8 records have high income.

$$P_{\text{low income}} = \frac{12}{20},\ P_{\text{high income}} = \frac{8}{20}$$

Entropy for low income

12 records (6 good, 6 bad)

$$-\frac{6}{12}\log_2\left(\frac{6}{12}\right) - \frac{6}{12}\log_2\left(\frac{6}{12}\right)$$

$$= -0.5\log_2(0.5) - 0.5\log_2(0.5)$$

$$= -0.5(-1) - 0.5(-1) = 1$$

Entropy for high income

8 records (4 good, 4 bad)

$$-\frac{4}{8}\log_2\left(\frac{4}{8}\right) - \frac{4}{8}\log_2\left(\frac{4}{8}\right)$$

$$= -0.5\log_2(0.5) - 0.5\log_2(0.5)$$

$$= -0.5(-1) - 0.5(-1) = 1$$

Combine the entropies of these two subsets,

$$\frac{12}{20}(1) + \frac{8}{20}(1)$$

$$= 0.6(1) + 0.4(1) = 1$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

1 – 1 = **0 bit**

Table 4.28 summarizes the information gain for each candidate split at decision node 13. Candidate split 1, Size of family has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.15

Table 4.28 Information Gain for each candidate split at decision node 13

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.03 bits |
| 2 | Parent's education level = Uneducated | 0.00 bits |
| 3 | Parent's employment status = Employed | 0.00 bits |
| 4 | Parents marital status = Married | 0.00 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.00 bits |

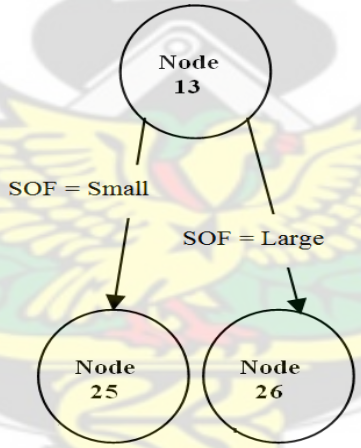


Figure 4.15 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of two new decision nodes. The 10 records at decision node 25 (Size of family = small) contain both good and bad academic performance. In the same vein, the 10 records at decision node 26 (Size of family = large) also contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.15 Candidate split at decision node 14

Determining the optimal split for decision node 13, containing records (11, 20, 35, 41, 45, 47, 72, 74, 77, 85, 87, 89, 91) as indicated in Table 4.29 below.

Table 4.29 Training set of records available at decision node 14

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **11** | Small | Uneducated | Unemployed | Married | Low | Good |
| **20** | Small | Uneducated | Unemployed | Married | High | Good |
| **35** | Small | Uneducated | Unemployed | Married | High | Bad |
| **41** | Large | Uneducated | Unemployed | Married | Low | Good |
| **45** | Large | Uneducated | Unemployed | Married | Low | Good |
| **47** | Small | Uneducated | Unemployed | Married | High | Good |
| **72** | Large | Uneducated | Unemployed | Married | Low | Good |
| **74** | Small | Uneducated | Unemployed | Married | High | Good |
| **77** | Small | Uneducated | Unemployed | Married | High | Good |
| **85** | Small | Uneducated | Unemployed | Married | High | Good |
| **87** | Large | Uneducated | Unemployed | Married | High | Good |
| **89** | Small | Uneducated | Unemployed | Married | Low | Bad |
| **91** | Small | Uneducated | Unemployed | Married | High | Bad |



Entropy before splitting

10 of the 13 records are classified as good academic performance, with the remaining 3 records classified as bad academic performance, the entropy before splitting is;

$$\text{PGAP} = \frac{10}{13}, \text{PBAP} = \frac{3}{13}$$

$$H(T) = -\sum_j P_j \log_2 (P_j)$$

$$-\frac{10}{13}\log_2\left(\frac{10}{13}\right) - \frac{3}{13}\log_2\left(\frac{3}{13}\right)$$

$$= -0.769 \log_2(0.769) - 0.231 \log_2(0.231)$$

$$= -0.769\,(-0.378) - 0.231\,(-2.114) = 0.779$$

After the analysis, compare the entropy of each candidate split against H(T) = 0.779 to see which split results in the greatest reduction in entropy (or gain in information).

**Size of family**

For candidate split 1 (size of family), 9 records have 'Small family' size and 4 records have 'Large family' size.

$$P_{small} = \frac{9}{13}, P_{large} = \frac{4}{13}$$

Entropy for small family size

9 records (6 good, 3 bad)

$$-\frac{6}{9}\log_2\left(\frac{6}{9}\right) - \frac{3}{9}\log_2\left(\frac{3}{9}\right)$$

$$= -0.667 \log_2(0.667) - 0.333 \log_2(0.333)$$

$$= -0.667\,(-0.584) - 0.333\,(-1.586) = 0.917$$

Entropy for large family size

4 records (4 good, 0 bad)

$$-\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right)$$

$$= -1\log_2(1) - 0\log_2(0)$$

$$= -1(0) - 0(0) = 0$$

Combine the entropies of these two subsets

$$\frac{9}{13}(0.917) + \frac{4}{13}(0)$$

$$= 0.692\ (0.917) + 0.307\ (0) = 0.634$$

The information gain represented by the split on the size of family attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

0.779 - 0.634 = **0.145 bit**

**Parent's education level**

For candidate split 2 (Parent's education level), all 13 records have 'uneducated' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parent's employment status**

For candidate split 3 (Parent's employment status), all 13 records have 'unemployed' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents marital status**

For candidate split 4 (Parents marital status), all 13 records have 'married' status. This will result in this attribute having the same number of bits as the entropy before splitting since they all use the same number of good and bad academic performance in the test set to perform the split. As a result, the entropy for this attribute is **0.000 bit**

**Parents monthly income**

For candidate split 5 (Parents monthly income), 5 records have low income and 8 records have high income.

$$P_{\text{low income}} = \frac{5}{13},\ P_{\text{high income}} = \frac{8}{13}$$

Entropy for low income

5 records (4 good, 1 bad)

$$-\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)$$

$$= -0.8\log_2(0.8) - 0.2\log_2(0.2)$$

$$= -0.8(-0.321) - 0.2(-2.321) = 0.721$$

Entropy for high income

8 records (6 good, 2 bad)

$$-\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)$$

$$= -0.75\log_2(0.75) - 0.25\log_2(0.25)$$

$$= -0.75(-0.415) - 0.25(-2) = 0.811$$

Combine the entropies of these two subsets,

$$\frac{5}{13}(0.721) + \frac{8}{13}(0.811)$$

$$= 0.384(0.721) + 0.615(0.811) = 0.775$$

The information gain represented by the split on the Parents monthly income attribute is;

$$\text{gain}(S) = H(T) - H_S(T)$$

$$0.779 - 0.775 = \textbf{0.004 bit}$$

Table 4.30 summarizes the information gain for each candidate split at decision node 14. Candidate split 1, Size of family has the largest information gain, and so is chosen for the initial split by the C4.5 algorithm. The partial decision tree resulting from C4.5's initial split is shown in Figure 4.16

Table 4.30 Information Gain for each candidate split at decision node 14

| Candidate Split | Child Nodes | Information Gain (Entropy Reduction) |
|---|---|---|
| 1 | Size of family = Small<br>Size of family = Large | 0.145 bits |
| 2 | Parent's education level = Uneducated | 0.000 bits |
| 3 | Parent's employment status = Unemployed | 0.000 bits |
| 4 | Parents marital status = Married | 0.000 bits |
| 5 | Monthly income = Low<br>Monthly Income = High | 0.004 bits |

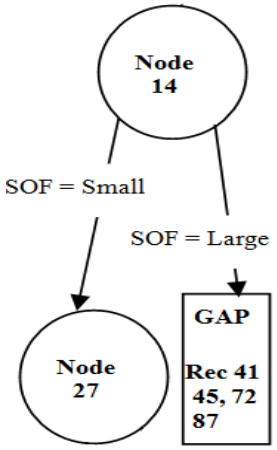


Figure 4.16 shows the partial decision tree resulting from C4.5's initial split.

The initial split has resulted in the creation of one terminal leaf node and one new decision node. Records with (Size of family = large) have only good academic performance therefore no further splits are required. The 9 records at decision node 27 (Size of family = small) contain both good and bad academic performance necessitating the need for further splitting.

### 4.2.16 Candidate split at decision node 15

Determining the optimal split for decision node 15, containing records (23, 28, 63) as indicated in Table 4.31 below.

Table 4.31 Training set of records available at decision node 15

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **23** | Small | Uneducated | Employed | Single | High | Good |
| **28** | Small | Uneducated | Employed | Single | High | Bad |
| **63** | Small | Uneducated | Employed | Single | Low | Good |

Because all the 4 other attributes are already splittered, monthly income automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with monthly income = low and monthly income = high have good and bad academic performance respectively.

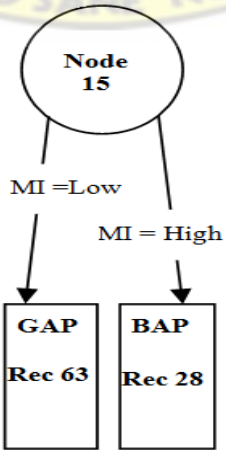


Figure 4.17 shows the partial decision tree resulting from C4.5's final split.

### 4.2.17 Candidate split at decision node 16

Determining the optimal split for decision node 16, containing records (46, 60, 86, 88) as indicated in Table 4.32 below.

Table 4.32 Training set of records available at decision node 16

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **46** | Small | Educated | Employed | Single | High | Bad |
| **60** | Small | Educated | Employed | Single | High | Bad |
| **86** | Small | Educated | Employed | Single | Low | Bad |
| **88** | Small | Educated | Employed | Single | Low | Bad |

Because all the 4 other attributes are already splittered, monthly income automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with monthly income = low and monthly income = high have good and bad academic performance respectively.

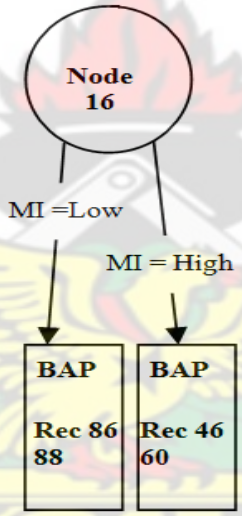


Figure 4.18 shows the partial decision tree resulting from C4.5's final split.

### 4.2.18 Candidate split at decision node 17

Determining the optimal split for decision node 17, containing records (10, 12, 19, 34, 40, 44, 54, 71, 73, 75, 80, 90) as indicated in Table 4.33 below.

Table 4.33 Training set of records available at decision node 17

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **10** | Large | Educated | Employed | Single | High | Bad |
| **12** | Large | Educated | Employed | Single | High | Bad |
| **19** | Large | Educated | Employed | Single | High | Bad |
| **34** | Large | Educated | Employed | Single | Low | Good |
| **40** | Large | Educated | Employed | Single | High | Good |
| **44** | Large | Educated | Employed | Single | Low | Bad |
| **54** | Large | Educated | Employed | Single | High | Good |
| **71** | Large | Educated | Employed | Single | High | Bad |
| **73** | Large | Educated | Employed | Single | Low | Bad |
| **75** | Large | Educated | Employed | Single | Low | Bad |
| **80** | Large | Educated | Employed | Single | Low | Good |
| **90** | Large | Educated | Employed | Single | Low | Good |

Because all the 4 other attributes are already splittered, monthly income automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with monthly income = low and monthly income = high have good and bad academic performance respectively.

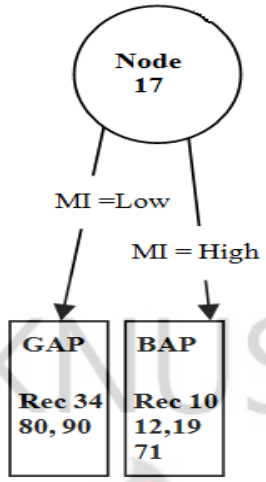


Figure 4.19 shows the partial decision tree resulting from C4.5's final split.

### 4.2.19 Candidate split at decision node 18

Determining the optimal split for decision node 18, containing records (4, 8, 17, 22, 36, 68, 78, 93, 95, 99) as indicated in Table 4.34 below.

Table 4.34 Training set of records available at decision node 18

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **4** | Large | Educated | Unemployed | Single | Low | Good |
| **8** | Large | Educated | Unemployed | Single | High | Bad |
| **17** | Large | Educated | Unemployed | Single | High | Bad |
| **22** | Large | Educated | Unemployed | Single | Low | Bad |
| **36** | Large | Educated | Unemployed | Single | Low | Good |
| **68** | Large | Educated | Unemployed | Single | Low | Bad |
| **78** | Large | Educated | Unemployed | Single | Low | Bad |
| **93** | Large | Educated | Unemployed | Single | Low | Good |
| **95** | Large | Educated | Unemployed | Single | High | Bad |
| **99** | Large | Educated | Unemployed | Single | Low | Bad |

Because all the 4 other attributes are already splittered, monthly income automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with monthly income = low and monthly income = high have good and bad academic performance respectively.

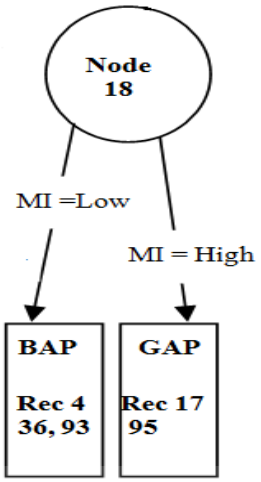


Figure 4.20 shows the partial decision tree resulting from C4.5's final split.



### 4.2.20 Candidate split at decision node 19

Determining the optimal split for decision node 19, containing records (52, 66) as indicated in Table 4.35 below.

Table 4.35 Training set of records available at decision node 19

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **52** | Small | Educated | Unemployed | Single | High | Good |
| **66** | Small | Educated | Unemployed | Single | Low | Bad |

Because all the 4 other attributes are already splittered, monthly income automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with monthly income = low and monthly income = high have good and bad academic performance respectively.

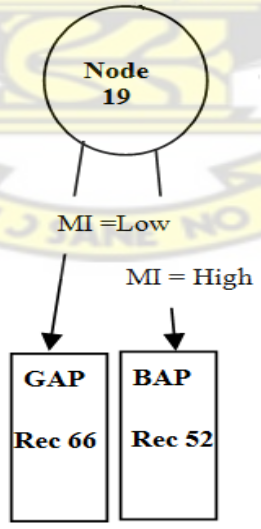


Figure 4.21 shows the partial decision tree resulting from C4.5's final split.

### 4.2.21 Candidate split at decision node 20

Determining the optimal split for decision node 20, containing records (2, 25, 30, 32, 42, 48, 97) as indicated in Table 4.36 below.

Table 4.36 Training set of records available at decision node 20

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| 2 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 25 | Large | Uneducated | Unemployed | Single | High | Good |
| 30 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 32 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 42 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 48 | Large | Uneducated | Unemployed | Single | Low | Bad |
| 97 | Large | Uneducated | Unemployed | Single | Low | Bad |

Because all the 4 other attributes are already splittered, monthly income automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with monthly income = low and monthly income = high have good and bad academic performance respectively.

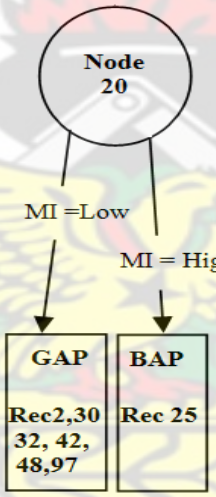


Figure 4.22 shows the partial decision tree resulting from C4.5's final split.

### 4.2.22 Candidate split at decision node 21

Determining the optimal split for decision node 21, containing records (38, 50, 83) as indicated in Table 4.37 below.

Table 4.37 Training set of records available at decision node 21

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| 38 | Small | Uneducated | Unemployed | Single | Low | Good |
| 50 | Small | Uneducated | Unemployed | Single | High | Bad |
| 83 | Small | Uneducated | Unemployed | Single | High | Bad |

Because all the 4 other attributes are already splittered, monthly income automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with monthly income = low and monthly income = high have good and bad academic performance respectively.

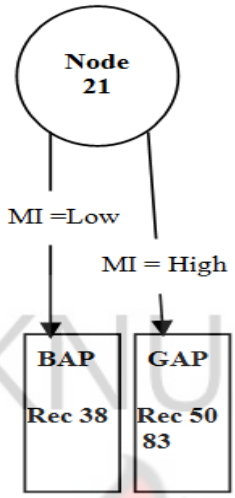


Figure 4.23 shows the partial decision tree resulting from C4.5's final split.

### 4.2.23 Candidate split at decision node 22

Determining the optimal split for decision node 22, containing records (6, 9, 24, 29, 49, 51, 84, 96) as indicated in Table 4.38 below.

Table 4.38 Training set of records available at decision node 22

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **6** | Large | Educated | Employed | Married | Low | Good |
| **9** | Large | Educated | Employed | Married | Low | Good |
| **24** | Small | Educated | Employed | Married | Low | Bad |
| **29** | Large | Educated | Employed | Married | Low | Bad |
| **49** | Large | Educated | Employed | Married | Low | Good |
| **51** | Large | Educated | Employed | Married | Low | Good |
| **84** | Large | Educated | Employed | Married | Low | Bad |
| **96** | Small | Educated | Employed | Married | Low | Good |

Because all the 4 other attributes are already splittered, size of family automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with small of family = small and size of family = large have good and bad academic performance respectively.

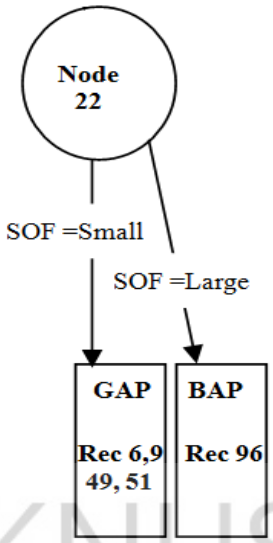


Figure 4.24 shows the partial decision tree resulting from C4.5's final split.

### 4.2.24 Candidate split at decision node 23

Determining the optimal split for decision node 23, containing records (13, 15, 62, 64, 81) as indicated in Table 4.39 below.

Table 4.39 Training set of records available at decision node 23

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **13** | Small | Educated | Unemployed | Married | High | Bad |
| **15** | Small | Educated | Unemployed | Married | High | Good |
| **62** | Large | Educated | Unemployed | Married | High | Bad |
| **64** | Large | Educated | Unemployed | Married | High | Bad |
| **81** | Small | Educated | Unemployed | Married | High | Bad |

Because all the 4 other attributes are already splittered, size of family automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with small of family = small and size of family = large have good and bad academic performance respectively.

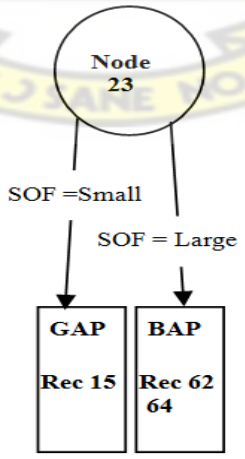


Figure 4.25 shows the partial decision tree resulting from C4.5's final split.

### 4.2.25 Candidate split at decision node 24

Determining the optimal split for decision node 24, containing records (27, 55, 58, 70) as indicated in Table 4.40 below.

Table 4.40 Training set of records available at decision node 24

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **27** | Large | Educated | Unemployed | Married | Low | Good |
| **55** | Small | Educated | Unemployed | Married | Low | Bad |
| **58** | Large | Educated | Unemployed | Married | Low | Good |
| **70** | Large | Educated | Unemployed | Married | Low | Bad |

Because all the 4 other attributes are already splittered, size of family automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with small of family = small and size of family = large have good and bad academic performance respectively.

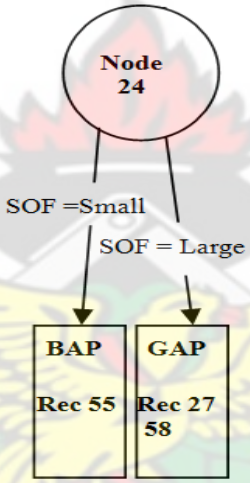


Figure 4.26 shows the partial decision tree resulting from C4.5's final split.

### 4.2.26 Candidate split at decision node 25

Determining the optimal split for decision node 25, containing records (3, 16, 21, 39, 53, 56, 59, 65, 79, 82) as indicated in Table 4.41 below.

Table 4.41 Training set of records available at decision node 25

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **3** | Large | Uneducated | Employed | Married | High | Bad |
| **16** | Large | Uneducated | Employed | Married | Low | Good |
| **21** | Large | Uneducated | Employed | Married | Low | Good |
| **39** | Large | Uneducated | Employed | Married | High | Bad |
| **53** | Large | Uneducated | Employed | Married | Low | Bad |
| **56** | Large | Uneducated | Employed | Married | High | Good |
| **59** | Large | Uneducated | Employed | Married | High | Good |
| **65** | Large | Uneducated | Employed | Married | Low | Good |
| **79** | Large | Uneducated | Employed | Married | High | Bad |
| **82** | Large | Uneducated | Employed | Married | Low | Good |

Because all the 4 other attributes are already splittered, monthly income automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with monthly income = low and monthly income = high have good and bad academic performance respectively.

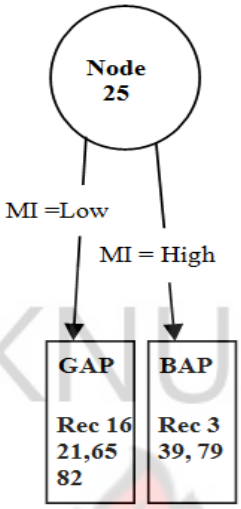


Figure 4.27 shows the partial decision tree resulting from C4.5's final split.

### 4.2.27 Candidate split at decision node 26

Determining the optimal split for decision node 26, containing records (5, 7, 18, 57, 67, 69, 76, 92, 94, 100) as indicated in Table 4.42 below.

Table 4.42 Training set of records available at decision node 26

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **5** | Small | Uneducated | Employed | Married | High | Bad |
| **7** | Small | Uneducated | Employed | Married | Low | Good |
| **18** | Small | Uneducated | Employed | Married | Low | Bad |
| **57** | Small | Uneducated | Employed | Married | Low | Bad |
| **67** | Small | Uneducated | Employed | Married | High | Good |
| **69** | Small | Uneducated | Employed | Married | High | Good |
| **76** | Small | Uneducated | Employed | Married | Low | Bad |
| **92** | Small | Uneducated | Employed | Married | Low | Good |
| **94** | Small | Uneducated | Employed | Married | Low | Bad |
| **100** | Small | Uneducated | Employed | Married | Low | Bad |

Because all the 4 other attributes are already splittered, monthly income automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with monthly income = low and monthly income = high have good and bad academic performance respectively.

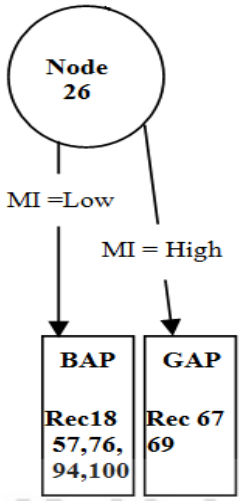


Figure 4.28 shows the partial decision tree resulting from C4.5's final split.

### 4.2.28 Candidate split at decision node 27

Determining the optimal split for decision node 27, containing records (11, 20, 35, 47, 74, 77, 85, 89, 91) as indicated in Table 4.43 below.

Table 4.43 Training set of records available at decision node 27

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| **11** | Small | Uneducated | Unemployed | Married | Low | Good |
| **20** | Small | Uneducated | Unemployed | Married | High | Good |
| **35** | Small | Uneducated | Unemployed | Married | High | Bad |
| **47** | Small | Uneducated | Unemployed | Married | High | Good |
| **74** | Small | Uneducated | Unemployed | Married | High | Good |
| **77** | Small | Uneducated | Unemployed | Married | High | Good |
| **85** | Small | Uneducated | Unemployed | Married | High | Good |
| **89** | Small | Uneducated | Unemployed | Married | Low | Bad |
| **91** | Small | Uneducated | Unemployed | Married | High | Bad |

Because all the 4 other attributes are already splittered, monthly income automatically performs the last split. This split has resulted in the creation of two terminal leaf nodes since both records with monthly income = low and monthly income = high have good and bad academic performance respectively.

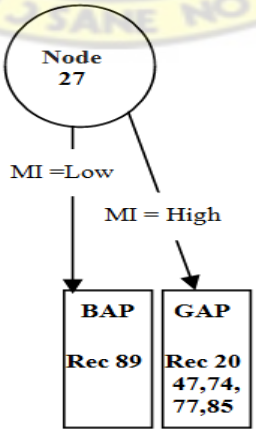


Figure 4.29 shows the partial decision tree resulting from C4.5's final split.

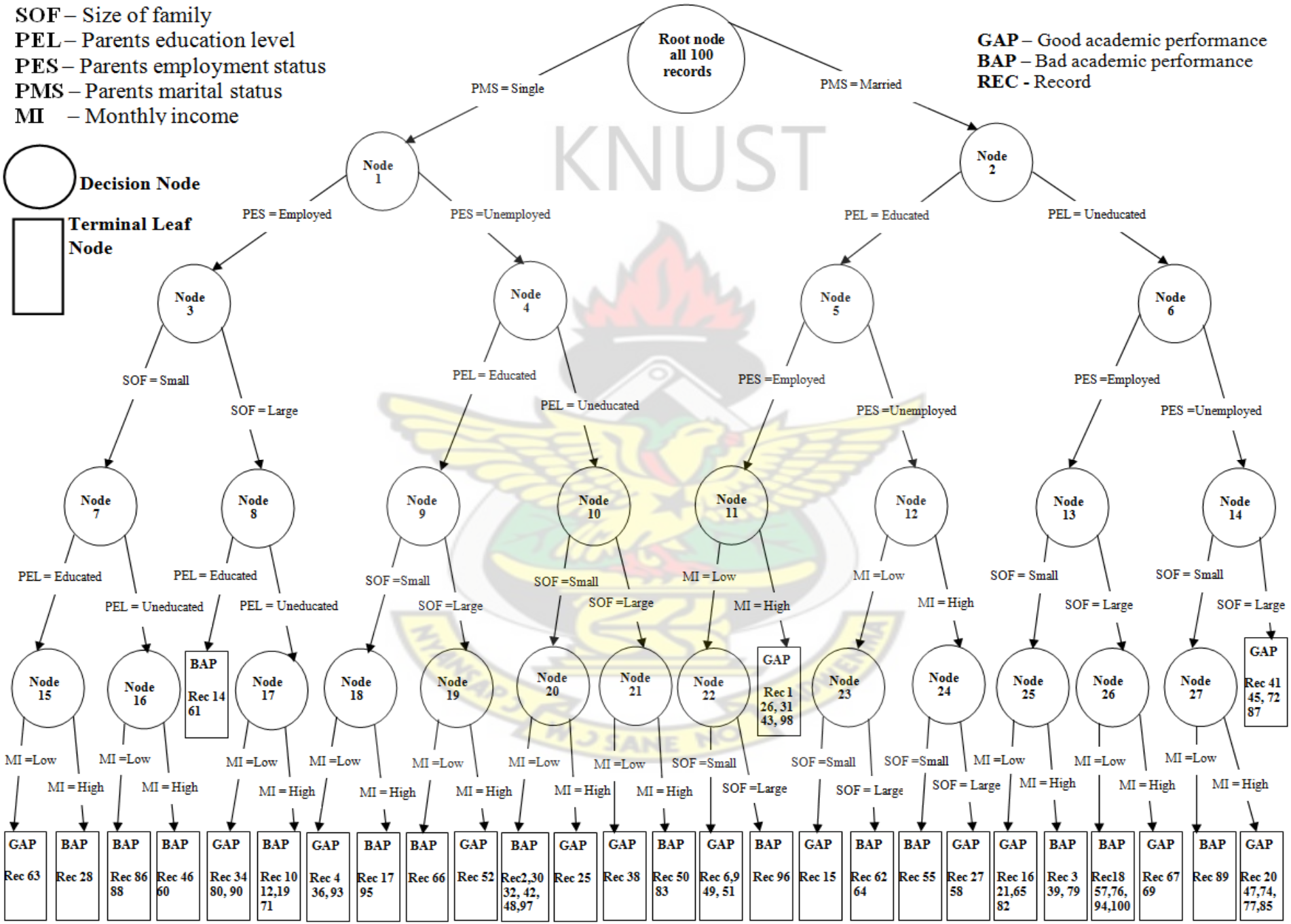


Figure 4.30 Academic Performance Classification Model (Decision Tree)

Table 4.44 Decision rules extracted from the decision tree in figure 4.30

| Antecedent | Consequent | Support | Confidence |
|---|---|---|---|
| If parents marital status = single and parents employment status = employed size of family = small and parents education level = educated and monthly income = low | Then good academic performance | $\frac{1}{100}$ | $\frac{1}{1} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = employed size of family = small and parents education level = educated and monthly income = high | Then bad academic performance | $\frac{1}{100}$ | $\frac{1}{1} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = employed size of family = small and parents education level = uneducated and monthly income = low | Then bad academic performance | $\frac{2}{100}$ | $\frac{2}{2} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = employed size of family = small and parents education level = uneducated and monthly income = high | Then bad academic performance | $\frac{2}{100}$ | $\frac{2}{2} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = employed size of family = large and parents education level = educated | Then bad academic performance | $\frac{2}{100}$ | $\frac{2}{2} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = employed size of family = large and parents education level = uneducated and monthly income = low | Then good academic performance | $\frac{3}{100}$ | $\frac{3}{3} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = employed size of family = large and parents education level = uneducated and monthly income = high | Then bad academic performance | $\frac{4}{100}$ | $\frac{4}{4} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = unemployed and parents education level = educated and size of family = small and monthly income = low | Then good academic performance | $\frac{3}{100}$ | $\frac{3}{3} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = unemployed and parents education level = educated and size of family = small and monthly income = high | Then bad academic performance | $\frac{2}{100}$ | $\frac{2}{2} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = unemployed and parents education level = educated and size of family = large and monthly income = low | Then bad academic performance | $\frac{1}{100}$ | $\frac{1}{1} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = unemployed and parents education level = educated and size of family = large and monthly income = high | Then good academic performance | $\frac{1}{100}$ | $\frac{1}{1} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = unemployed and parents education level = uneducated and size of family = small and monthly income = low | Then bad academic performance | $\frac{6}{100}$ | $\frac{6}{6} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = unemployed and parents education level = uneducated and size of family = small and monthly income = high | Then good academic performance | $\frac{1}{100}$ | $\frac{1}{1} * 100 = 100\%$ |
| If parents marital status = single and parents employment status = unemployed and parents education level = uneducated and size of family = large and monthly income = low | Then good academic performance | $\frac{1}{100}$ | $\frac{1}{1} * 100 = 100\%$ |

| | | | |
|---|---|---|---|
| If parents marital status = single and parents employment status = unemployed and parents education level = uneducated and size of family = large and monthly income = high | Then bad academic performance | $\frac{2}{100}$ | $\frac{2}{2} * 100 = 100\%$ |
| If parents marital status = married and parents education level = educated and parents employment status = employed and monthly income = low and size of family = small | Then good academic performance | $\frac{4}{100}$ | $\frac{4}{4} * 100 = 100\%$ |
| If parents marital status = married and parents education level = educated and parents employment status = employed and monthly income = low and size of family = large | Then bad academic performance | $\frac{1}{100}$ | $\frac{1}{1} * 100 = 100\%$ |
| If parents marital status = married and parents education level = educated and parents employment status = employed and monthly income = high | Then good academic performance | $\frac{5}{100}$ | $\frac{5}{5} * 100 = 100\%$ |
| If parents marital status = married and parents education level = educated and parents employment status = unemployed and monthly income = low and size of family = small | Then good academic performance | $\frac{1}{100}$ | $\frac{1}{1} * 100 = 100\%$ |
| If parents marital status = married and parents education level = educated and parents employment status = unemployed and monthly income = low and size of family = large | Then bad academic performance | $\frac{2}{100}$ | $\frac{2}{2} * 100 = 100\%$ |
| If parents marital status = married and parents education level = educated and parents employment status = unemployed and monthly income = high and size of family = small | Then bad academic performance | $\frac{1}{100}$ | $\frac{1}{1} * 100 = 100\%$ |
| If parents marital status = married and parents education level = educated and parents employment status = unemployed and monthly income = high and size of family = large | Then good academic performance | $\frac{2}{100}$ | $\frac{2}{2} * 100 = 100\%$ |
| If parents marital status = married and parents education level = uneducated and parents employment status = employed and size of family = small monthly income = low | Then good academic performance | $\frac{4}{100}$ | $\frac{4}{4} * 100 = 100\%$ |
| If parents marital status = married and parents education level = uneducated and parents employment status = employed and size of family = small monthly income = high | Then bad academic performance | $\frac{3}{100}$ | $\frac{3}{3} * 100 = 100\%$ |
| If parents marital status = married and parents education level = uneducated and parents employment status = employed and size of family = large and monthly income = low | Then bad academic performance | $\frac{5}{100}$ | $\frac{5}{5} * 100 = 100\%$ |
| If parents marital status = married and parents education level = uneducated and parents employment status = employed and size of family = large and monthly income = high | Then good academic performance | $\frac{2}{100}$ | $\frac{2}{2} * 100 = 100\%$ |
| If parents marital status = married and parents education level = uneducated and parents employment status = unemployed and size of family = small and monthly income = low | Then bad academic performance | $\frac{1}{100}$ | $\frac{1}{1} * 100 = 100\%$ |
| If parents marital status = married and parents education level = uneducated and parents employment status = unemployed and size of family = small and monthly income = high | Then good academic performance | $\frac{5}{100}$ | $\frac{5}{5} * 100 = 100\%$ |
| If parents marital status = married and parents education level = uneducated and parents employment status = unemployed and size of family = large | Then good academic performance | $\frac{4}{100}$ | $\frac{4}{4} * 100 = 100\%$ |

### 4.2.29 Decision rule interpretation

One of the most attractive aspects of decision trees lies in their interpretability, especially with respect to the construction of decision rules (Larose, 2005). Rules are a good way of representing information or bits of knowledge. The complete set of decision rules generated by a decision tree is equivalent (for classification purposes) to the decision tree itself (Larose, 2005). The rules come in the form; if antecedent, then consequent. For decision rules, the antecedent consists of the attribute values from the branches taken by the particular path through the tree, while the consequent consists of the classification value for the target variable given by the particular leaf node. (Larose, 2005). A rule-based classifier also uses a set of IF-THEN rules for classification, (Han & Kamber, 2006). An IF-THEN rule is an expression of the form; IF condition THEN conclusion. The IF part (or left-hand side) of a rule is the rule antecedent or precondition. The THEN part (or right-hand side) is the rule consequent. In the rule antecedent, the condition consists of one or more attribute tests and the rule's consequent contains a class prediction. According to Han & Kamber, 2006, rules extracted directly from the tree are mutually exclusive and exhaustive. This means that we cannot have rule conflicts because no two rules will be triggered for the same tuple. We have one rule per leaf, and any tuple can map to only one leaf.

### 4.2.30 Confidence and Support

According to (Larose, 2005), the support of the decision rule refers to the proportion of records in the data set that rest in that particular terminal leaf node whiles the confidence of the rule refers to the proportion of records in the leaf node for which the decision rule is true. In other words, out of the number of records that support the rule, how many support good or bad academic performance proportionately in the data set for which the decision rule is true? In this study, all the leaf nodes are pure, resulting in perfect confidence levels of 100%. From the generated decision tree in Figure 4.30, the rules above were extracted to serve as the hypotheses for the training data to be tested on.

## 4.3 Test set

In a real-world application of supervised learning, we have a training set of examples with labels, and a test set of examples with unknown labels. The whole point is to make predictions for the test examples. (Elkan, 2012). However, in research, we want to measure the performance achieved by a learning algorithm and to do this we use a test set consisting of examples with known labels, (Elkan, 2012). We train the classifier on the training set, apply it to the test set, and then measure performance by comparing the predicted labels with the true labels. A common rule of thumb is to use 70% of the

database for training and 30% for testing, (Elkan, 2012). It is absolutely vital to measure the performance of a classifier on an independent test set. Every training algorithm looks for patterns in the training data, i.e. correlations between the features and the class. Only accuracy measured on an independent test set is a fair estimate of accuracy on the whole population, (Elkan, 2012). The phenomenon of relying on patterns that are strong only in the training data is called overfitting. In practice it is an omnipresent danger, (Elkan, 2012). The following test data is used to measure how accurate the model performs on unseen data.

### 4.3.1 Candidate split at root node

Determining the optimal split for the root node containing records (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50) as indicated in Table 4.45 below.

Table 4.45 Test set of records at the root node

| Student | Size of family | Parents education level | Parents employment status | Parents marital status | Monthly income | Academic Performance |
|---|---|---|---|---|---|---|
| 1 | Large | Educated | Employed | Married | Low | Good |
| 2 | Large | Educated | Unemployed | Single | Low | Bad |
| 3 | Small | Uneducated | Employed | Married | High | Good |
| 4 | Small | Educated | Employed | Married | Low | Bad |
| 5 | Large | Uneducated | Unemployed | Single | High | Good |
| 6 | Small | Educated | Employed | Married | High | Good |
| 7 | Large | Educated | Unemployed | Married | Low | Good |
| 8 | Small | Uneducated | Employed | Single | High | Bad |
| 9 | Large | Educated | Employed | Married | Low | Good |
| 10 | Large | Uneducated | Unemployed | Single | High | Bad |
| 11 | Small | Educated | Employed | Single | High | Bad |
| 12 | Small | Uneducated | Unemployed | Married | Low | Good |
| 13 | Large | Educated | Employed | Married | Low | Bad |
| 14 | Small | Uneducated | Unemployed | Married | High | Good |
| 15 | Large | Educated | Employed | Single | Low | Bad |
| 16 | Small | Educated | Employed | Married | Low | Bad |
| 17 | Small | Uneducated | Employed | Married | High | Good |
| 18 | Large | Educated | Unemployed | Single | Low | Bad |
| 19 | Small | Uneducated | Employed | Married | High | Bad |
| 20 | Large | Educated | Employed | Single | Low | Good |
| 21 | Large | Educated | Employed | Married | Low | Good |
| 22 | Small | Educated | Unemployed | Single | High | Good |
| 23 | Large | Uneducated | Employed | Married | Low | Bad |
| 24 | Small | Educated | Employed | Single | High | Bad |
| 25 | Small | Educated | Unemployed | Married | Low | Good |
| 26 | Large | Uneducated | Employed | Single | High | Good |
| 27 | Small | Educated | Employed | Married | Low | Bad |
| 28 | Large | Educated | Unemployed | Single | High | Good |
| 29 | Large | Uneducated | Employed | Married | High | Good |
| 30 | Small | Educated | Employed | Single | low | Good |
| 31 | Small | Uneducated | Unemployed | Married | Low | Good |
| 32 | Large | Educated | Employed | Single | Low | Good |
| 33 | Small | Educated | Unemployed | Married | High | Bad |
| 34 | Large | Uneducated | Employed | Single | Low | Good |
| 35 | Small | Educated | Unemployed | Married | High | Bad |
| 36 | Large | Uneducated | Employed | Married | Low | Bad |
| 37 | Large | Educated | Unemployed | Single | High | Bad |
| 38 | Small | Uneducated | Employed | Married | Low | Bad |
| 39 | Large | Educated | Employed | Single | High | Good |
| 40 | Small | Uneducated | Unemployed | Married | High | Good |
| 41 | Small | Educated | Employed | Married | High | Good |
| 42 | Large | Educated | Employed | Single | Low | Good |
| 43 | Small | Uneducated | Employed | Married | High | Bad |
| 44 | Large | Educated | Unemployed | Single | Low | Good |
| 45 | Small | Uneducated | Employed | Married | High | Bad |
| 46 | Large | Educated | Employed | Single | Low | Good |
| 47 | Small | Uneducated | Employed | Married | High | Good |
| 48 | Large | Educated | Unemployed | Single | High | Bad |
| 49 | Small | Educated | Employed | Married | Low | Good |
| 50 | Small | Educated | Employed | Single | High | Bad |

The test data above is used to produce the decision tree below for the test model after performing the data mining analysis. The analysis is not represented here due to the voluminous nature of the document.

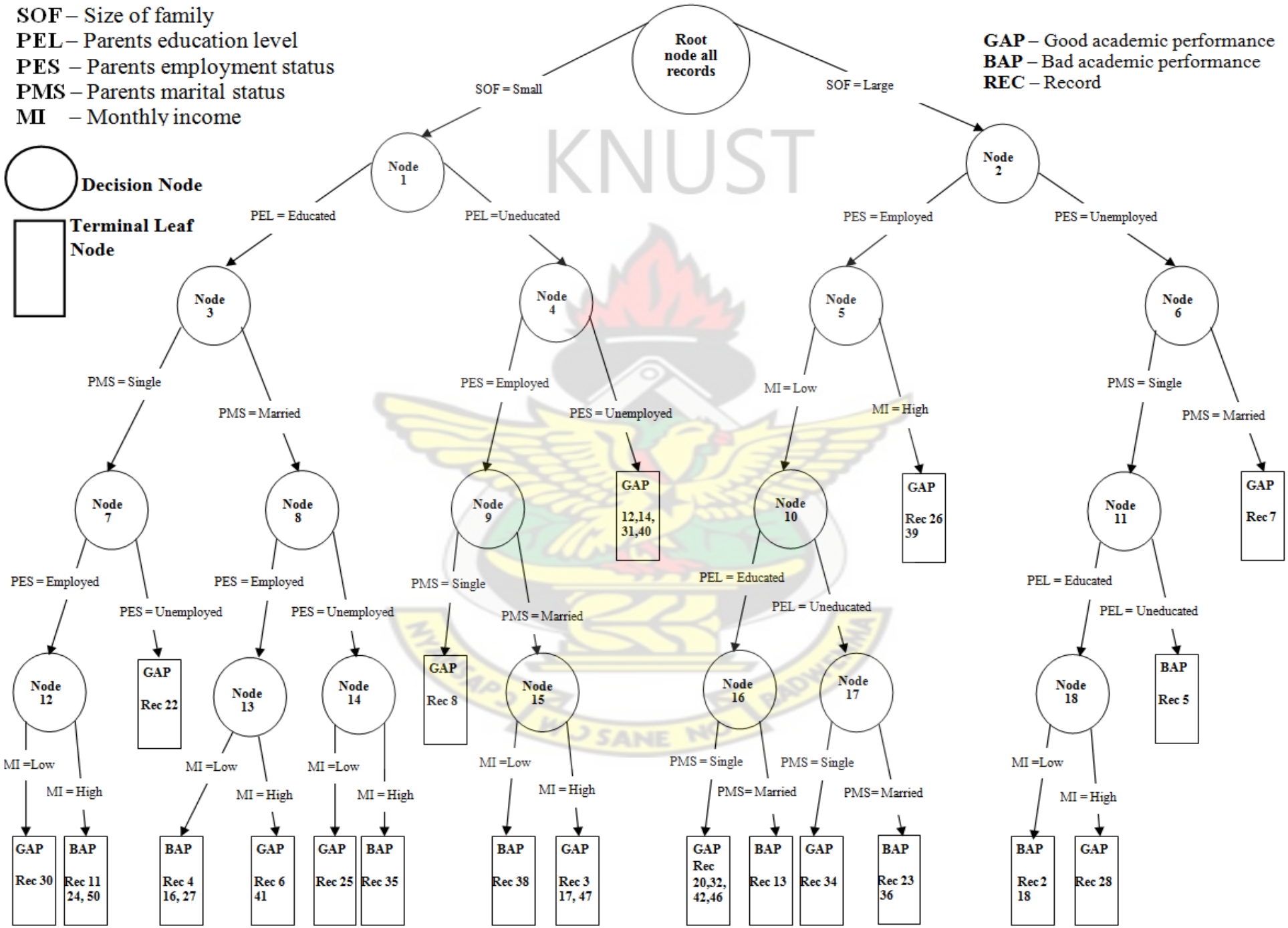


Figure 4.50 Test Model (Decision Tree)

### 4.3.20 Decision rule coverage

A rule's coverage is the percentage of tuples that are covered by the rule (i.e., whose attribute values hold true for the rule's antecedent). For a rule's accuracy, we look at the tuples that it covers and see what percentage of them the rule can correctly classify, (Han & Kamber, 2006). The following rules were extracted and the rule set produced were the following;

Table 4.76 Decision rules extracted from the decision tree in figure 4.50

| Antecedent | Consequent | Support | Confidence |
|---|---|---|---|
| If size of family = small and parents education level = educated and parents marital status = single and parents employment status = employed and monthly income = low | Then good academic performance | $\frac{1}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = small and parents education level = educated and parents marital status = single and parents employment status = employed and monthly income = high | Then bad academic performance | $\frac{3}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = small and parents education level = educated and parents marital status = single and parents employment status = unemployed | Then good academic performance | $\frac{1}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = small and parents education level = educated and parents marital status = married and parents employment status = employed and monthly income = low | Then bad academic performance | $\frac{3}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = small and parents education level = educated and parents marital status = married and parents employment status = employed and monthly income = high | Then good academic performance | $\frac{2}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = small and parents education level = educated and parents marital status = married and parents employment status = unemployed and monthly income = low | Then good academic performance | $\frac{1}{50}$ | $\frac{2}{2} * 100 = 100\%$ |
| If size of family = small and parents education level = educated and parents marital status = married and parents employment status = unemployed and monthly income = high | Then bad academic performance | $\frac{1}{50}$ | $\frac{3}{3} * 100 = 100\%$ |
| If size of family = small and parents education level = uneducated and parents employment status = employed parents marital status = single | Then good academic performance | $\frac{1}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = small and parents education level = uneducated and parents employment status = employed parents marital status = married monthly income = low | Then bad academic performance | $\frac{1}{50}$ | $\frac{1}{1} * 100 = 100\%$ |

| | | | |
|---|---|---|---|
| If size of family = small and parents education level = uneducated and parents employment status = employed parents marital status = single monthly income = high | Then good academic performance | $\frac{3}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = small and parents education level = uneducated and parents employment status = unemployed | Then good academic performance | $\frac{4}{50}$ | $\frac{4}{4} * 100 = 100\%$ |
| If size of family = large and parents employment status = employed and monthly income = low parents education level = educated and parents marital status = single | Then good academic performance | $\frac{4}{50}$ | $\frac{4}{4} * 100 = 100\%$ |
| If size of family = large and parents employment status = employed and monthly income = low parents education level = educated and parents marital status = married | Then bad academic performance | $\frac{1}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = large and parents employment status = employed and monthly income = low parents education level = uneducated and parents marital status = single | Then good academic performance | $\frac{1}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = large and parents employment status = employed and monthly income = low parents education level = uneducated and parents marital status = married | Then bad academic performance | $\frac{2}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = large and parents employment status = employed and monthly income = high | Then good academic performance | $\frac{2}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = large and parents employment status = unemployed and parents marital status = single and parents education level = educated and monthly income = low | Then bad academic performance | $\frac{2}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = large and parents employment status = unemployed and parents marital status = single and parents education level = educated and monthly income = high | Then good academic performance | $\frac{1}{50}$ | $\frac{2}{2} * 100 = 100\%$ |
| If size of family = large and parents employment status = unemployed and parents marital status = single and parents education level = uneducated | Then bad academic performance | $\frac{1}{50}$ | $\frac{1}{1} * 100 = 100\%$ |
| If size of family = large and parents employment status = unemployed and parents marital status = married | Then bad academic performance | $\frac{1}{50}$ | $\frac{1}{1} * 100 = 100\%$ |

The performance evaluation of this model was focused on the concepts of sensitivity, specificity and accuracy in the context of academic performance classification. Given the importance of these statistics in model evaluation, they are further discussed in the preceding sections.

### 4.4 Measures of classification success

According to Ian & Eibe (2005), 'the numerical error rate of a classifier on a test set may be in the region of 25%'. This corresponds to a success rate of 75%; representing the true success rate of the classifier on the target population. To predict the performance of a classifier on new data, we need to assess its error rate on a dataset that played no part in the formation of the classifier. This independent

dataset is called the test set, (Ian & Eibe, 2005). Provided both samples are representative, the error rate on the test set will give a true indication of future performance and the accuracy of the error estimate can be quantified statistically (Ian & Eibe, 2005). The evaluation of the performance of the classification model is based on the counts of test records correctly and incorrectly as predicted by the model, (Tan, 2005). Variables of academic performance include different kinds of information, such as family size, parents education level, peer pressure, monthly income, marital status, employment status, school attendance, etc. Schools' decisions on academic performance rely on factors that stimulate the academic performance of students, thus making the accuracy of classification models essential in education administration. For the purposes of this study, the best possible classifier on the given academic performance variables were chosen based on two attributes. The attributes are measured to determine how successful the model performs on an unseen data. For supervised learning with two possible classes, all measures of performance are based on four numbers obtained from applying the classifier to the test set, (Elkan, 2012). These numbers are called true positives TP, false positives FP, true negatives TN, and false negatives FN. Positive classification being good academic performance and negative classification being bad academic performance. These measures of performance are tabulated in a confusion matrix. The table below shows the confusion matrix for the academic performance classification on the test data.

Table 4.77 Sensitivity, specificity and accuracy on the test data

| **Outcome of the classification process** | **Academic Performance classification as determined by the standard of truth on the test data** | | |
|---|---|---|---|
| | **Positive** | **Negative** | **Row Total** |
| **Positive** | **TP (8)** | **FP (2)** | **TP+FP (10)** (Total number of subjects with positive classification) |
| **Negative** | **FN (3)** | **TN (6)** | **FN+TN (9)** (Total number of subjects with positive test) |
| **Column total** | **TP+FN (11)** | **FP+TN (8)** | **N = TP+TN+ FP+FN (19)** (Total number of subjects in study) |

### 4.5 True and False Positives and Negatives (Type I and Type II error)

The two types of classification errors are false positives and false negatives, (Bramer, 2007). False positives (Type 1 Errors) occur when instances that should be classified as negative are classified as positive. False negatives (Type 2 Errors) occur when instances that should be classified as positive are classified as negative, (Bramer, 2007). The values of P and N, the number of positive and negative instances, are fixed for a given test set irrespective of whichever classifier is used. Given the values of True Positive Rate and False Positive Rate as well as P and N, all the other measurements for the model were derived. This model is therefore characterized by its True Positive Rate (TP Rate) and False Positive Rate (FP Rate) values, which are both proportions from 0 to 1, (Bramer, 2007). If a student's academic performance is proven to be good and it is corroborated by the given test data, the result of the classification is considered true positive (TP). Similarly, if a student's academic performance is proven is proven to be bad and the test data suggests that, the test result is true negative (TN). Both true positive and true negative suggest a consistent result between the test data and the proven condition (Standard of truth), Zhu et al, (2010). However, no model is perfect; if the test data indicates the presence of bad academic performance of a student who actually belong to the good academic performance class, the test result is false negative (FN). Similarly, if the result of the test data suggests the presence of good academic performance of a student who actually belongs to the bad academic performance class, the test result is false positive (FP). Both false positive and false negative indicate that the test results are opposite to the actual condition. Sensitivity, specificity and accuracy are described in terms of TP, TN, FN and FP and the evaluation measure mostly used in practice is the accuracy rate (Acc). Zhu et al, (2010). According to them, 'Accuracy is the proportion of true results, either true positive or true negative, in a population'. The accuracy of the model was achieved by measuring the degree of veracity of the test data on the model. This also involved the evaluation of the effectiveness of the model by its percentage of correct predictions, Zhu et al, (2010). Equation 1 shows how Accuracy was computed on the test data. |A| denotes the cardinality of set A. (the measure of the number of elements of set A). Costa et al, (2007)

$$\text{Acc} = \frac{|\text{TN}| + |\text{TP}|}{|\text{FN}| + |\text{FP}| + |\text{TN}| + |\text{TP}|} \qquad \text{Equation 1}$$

$$\text{Acc} = \frac{8+6}{3+2+6+8} = \frac{14}{19} = 0.74$$

The complement of Accuracy (Acc) is the error rate (Err). Equation 2, evaluates the model by its percentage of incorrect predictions. Costa et al, (2007)

$$Err = \frac{|FN| + |FP|}{|FN| + |FP| + |TN| + |TP|} \quad \text{Equation 2}$$

$$Err = \frac{3+2}{3+2+6+8} = \frac{5}{19} = 0.26$$

The recall (R) measured and evaluated the effectiveness of the model for each class in the test data by applying the test data on the model. Recall (sensitivity or true positive rate) is the proportion of samples belonging to the positive class which were correctly predicted as positive by the model. Costa et al, (2007). It showed how good the academic performance classification model is at classifying a student belonging to the good academic performance class. Its numerical value represented the probability of the model to truly identify students who belong to this class. The higher the numerical value, the less likely for the model to return false positive results. Equation 3 shows how Recall (R) was computed on the test data, Costa et al, (2007).

$$R = \frac{|TP|}{|TP| + |FN} \quad \text{Equation 3}$$

$$R = \frac{8}{8+3} = \frac{8}{11} = 0.73$$

Specificity (Spe) measured and evaluated the effectiveness of the model for each class in the test data in comparison to the training data. Specificity (true negative rate) is the percentage of negative samples correctly predicted as negative by applying the test data on the model. Equation 4 shows how Specificity (Spe) was computed on the test data. Costa et al, (2007) & Zhu et al, (2010).

$$Spe = \frac{|TN|}{|FP| + |TN|} \quad \text{Equation 4}$$

$$Spe = \frac{6}{6+2} = \frac{6}{8} = 0.75$$

(1 – Specificity)

1 - 0.75 = 0.25

Precision (P) is measured to estimate the probability that the positive predictions were correct. Its computation is given by Equation 5. Costa et al, (2007).

$$P = \frac{|TP|}{|TP| + |FP|}$$ Equation 5

$$P = \frac{8}{8+2} = \frac{8}{10} = 0.8$$

Precision and Recall are combined into a single measure, (their product divided by their average to get the harmonic mean) which is the measurement of central location to aggregate the scores in the two measurements to determine their single final score, (Gupta & Malviya, 2012).

$$F - Measure = \frac{2 * Recall * Precision}{Recall + Precision}$$ Equation 6

$$F - Measure = \frac{2*0.73*0.8}{0.73+0.8} = \frac{5}{19} = 2.4$$

Based on the calculations of the equations above, the academic performance classification model has a sensitivity of 73%, this means that there is a 73%, chance that, the model will classify students who belong to the good academic performance class correctly, Zhu et al, (2010). The numerical value of specificity represented the probability of the test data to classify students without giving false positive results, Zhu et al, (2010). The academic performance classification model has a specificity of 75%. This means that when new data is thrown at the model to classify, there is a 75% chance that, the model will classify a student belonging to the bad academic performance class accurately, Zhu et al, (2010). According to them, 'a test can be very specific without being sensitive, or it can be very sensitive without being specific. Both factors are equally important but a good test is one that has both high sensitivity and specificity'. The numerical value of accuracy represents the proportion of positive results (both true positive and true negative) in the selected population Zhu et al, (2010). The academic performance classification model has an accuracy of 74%. This reflects the percentage results of how accurate the model is regardless of the positive or negative classification. However, it worth mentioning that, the equation of accuracy implies that even if both sensitivity and specificity are high, it does not suggest that the accuracy of the test data is equally high. In addition to sensitivity and specificity, the accuracy is also determined by how common the model classifies students from the good or bad academic performance class within the test data, Zhu et al, (2010).

### 4.6 Receiver operating characteristics (ROC) graph visualization

Another common evaluation measure used in binary classification problems is the ROC (Receiver

Operating Characteristics) graph, which relates sensitivity and specificity, (Fawcett, 2003). This graph is a technique for visualizing, organizing and selecting classifiers based on their performance and serves as a useful technique for organizing classifiers and visualizing their performance, (Fawcett, 2003). These graphs have long been used in signal detection theory to depict the tradeoff between hit rates and false alarm rates of classifiers (Fawcett, 2003). One of the earliest adopters of ROC graphs in machine learning was Spackman (1989), who demonstrated the value of ROC curves in evaluating and comparing algorithms. Recent years have seen an increase in the use of ROC graphs in the machine learning community, (Fawcett, 2003). For the academic performance classification model, the true positive rate (TPR) was matched against the false positive rate (FPR) to determine the single point in the ROC graph. This point showed the tradeoff between sensitivity and specificity, an increase in sensitivity is accompanied by a decrease in specificity, Zhu et al, (2010). From the equations above, TPR is equivalent to sensitivity and FPR is equivalent to (1 – specificity), Zhu et al, (2010). All possible combinations of TPR and FPR compose the ROC space. This single cut point of the model on the ROC graph defines one single point in the ROC space, Zhu et al, (2010). Thus the location of the point in the ROC space depicts whether the academic performance classification model was a good classifier or not, Zhu et al, (2010). The model had 73% sensitivity and (1 – 0.75) 25% specificity. In an ideal situation, the point determined by both TPR and FPR yields a coordinates (0, 1), and this point falls on the upper left corner of the ROC space. The cut point on the ROC space indicates that, the academic performance classification model has a sensitivity of 73% and specificity of 25% and this falls within the area above the diagonal line which represents a good academic performance classification, otherwise a bad prediction. The diagonal line joining the bottom left and top right-hand corners correspond to random guessing, whatever the probability of the positive class may be, (Bramer, 2007). Theoretically, a random guess would give a point along this diagonal, Zhu et al, (2010). The lower right-hand triangle corresponds to classifiers that are worse than random guessing, (Bramer, 2007). This can be visualized on the diagonal determined by coordinate (0, 73) on the y-axis and coordinates (0, 25) on the x-axis. Any classifier that appears in the lower right triangle performs worse than random guessing, (Fawcett, 2003). The interpretation of ROC curve is similar to a single point in the ROC space; the slope of the tangent line to a cut-point tells us the ratio of the probability of identifying true positive over true negative and the selected cut-point doesn't also add additional information to identify the true positive result, Zhu et al, (2010). A graphical view of the ROC analysis on the test data is

presented in the figure below.

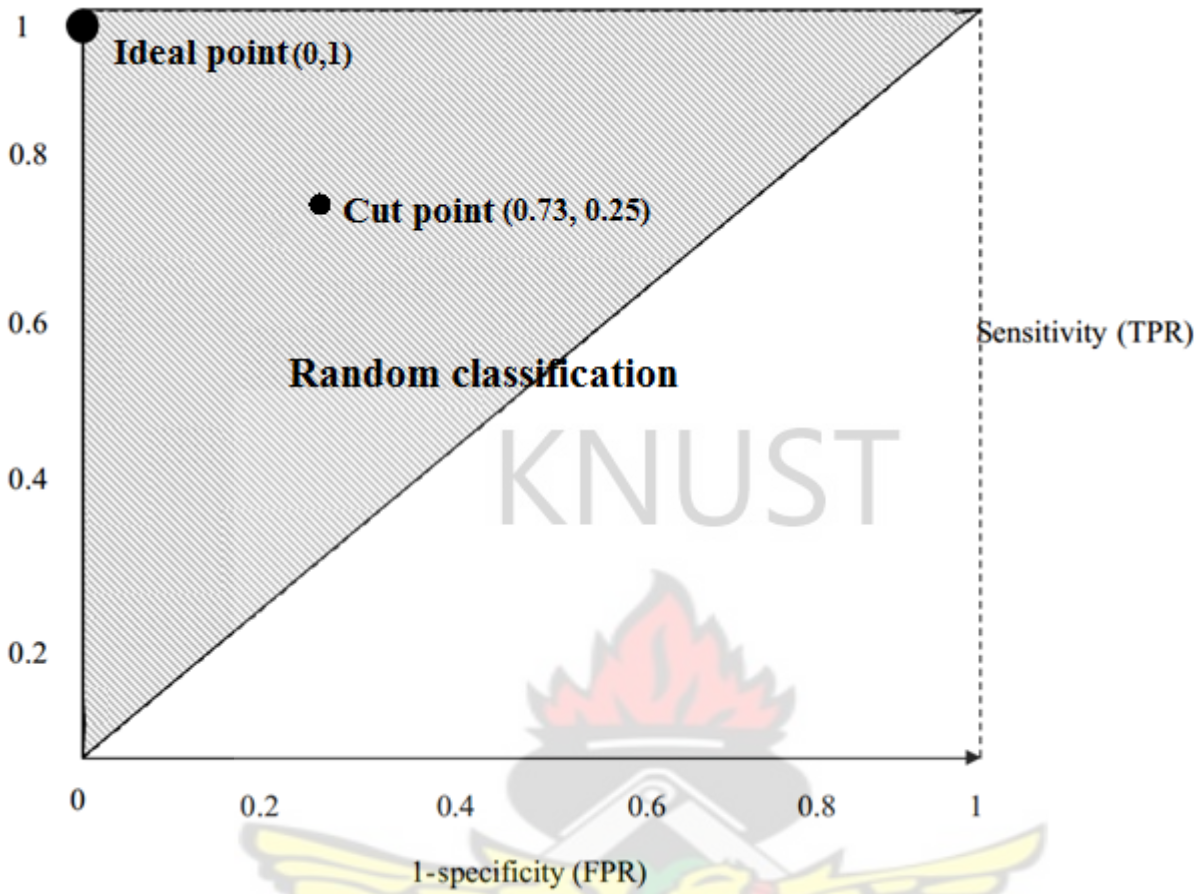


Figure 4.51 ROC graph: the shadow area represents a better model classification

# CHAPTER 5

# SUMMARY, CONCLUSION AND RECOMMENDATION

## 5.1 Summary

The top down induction of decision trees is one of the most commonly used methods of classification and it is widely cited in research literature, (Bramer, 2007). For classification problems, it is natural to measure a classifier's performance in terms of the error rate. The classifier predicts the class of each instance if it is counted as a success; otherwise, it is an error (Ian & Eibe, 2005). The error rate is just the proportion of errors made over a whole set of instances, and it measures the overall performance of the classifier. Ian and Eibe (2005) stated that, 'the resubstitution error is the error rate on the training data and it is calculated by resubstituting the training instances into a classifier that was constructed from them'. It is not a reliable predictor of the true error rate on new data. Sensitivity, specificity and accuracy which are widely used statistical methods in model evaluation are used to measure the correctness of the model, Zhu et al, (2010). Sensitivity evaluates how good the model is at detecting a student from the good academic performance class while Specificity estimates how likely students with bad academic performance will be detected Zhu et al, (2010). The accuracy of the model can therefore be determined from these two accuracy measures. The ROC graph is the graphic presentation of the relationship between sensitivity and specificity. It helps to decide on the models optimality through the determination of the best threshold for the classifier Zhu et al, (2010).

## 5.2 Conclusion and future work

Synthesizing the vast amount of research and ideas and condensing them with the aim of introducing data mining to the public basic education system is a great challenge. By using well defined algorithms from the disciplines of machine learning and artificial intelligence to discern rules, data mining has profound application significance, (Luan, 2000). In this study, I have shown how useful data mining can be used at the basic education level, particularly to help improve students' academic performance by accurately predicting student's academic performance based on certain socio-economic variables. The application of this model can be used to develop performance monitoring and evaluation systems. This model also has the potential to improve performance monitoring of public basic schools by offering historical perspectives of students' academic performance. Although the continuous evaluation of students' academic performance is among the principles of improving student learning, evidence of applying data mining to the socio-economic challenges of students' academic performance is lacking in Ghana. This is partly due to the lack of adequate data to undertake such research. This

study represents a step towards helping to fill this gap. I used questionnaire to collect data from 10 selected public basic schools in the Ablekuma west constituency of the Greater Accra region to determine whether the selected socio-economic variables can effectively help in the development of a prediction rule model to predict students' academic performance. I applied data mining techniques (C4.5 algorithm) to discover knowledge in this regard, particularly the decision tree. I then used several classification accuracy measures to assess the accuracy of the model. The list of socio-economic factors investigated could not have been exhaustive; there are several other factors that can influence academic performance. My future work include applying data mining techniques on an expanded data set covering a much wider population and with more distinctive attributes to get more accurate results. Also, experiments could be done using other data mining techniques such as neural nets, genetic algorithms and k-nearest Neighbor. The tools and techniques presented in this study provide a starting point for the development of classification models for schools in Ghana. Finally, the used preprocess that cumulated in the development of this model could be embedded into a software so that public basic schools can derive the maximum benefit from it. It's my hope that the results of this study will be useful to policymakers, researchers, and others interested in the application of data mining in educational research so as to complement and supplement basic education performance monitoring and assessment programs.

### 5.3 Recommendations

For the purpose of this study, the following intervention measures are recommended for adoption by the Ghana Education Service;

- Headteachers of the various public basic schools should not only focus their attention on classroom activities but also on the economic and social needs of their students.
- Schools should also devise means of paying special attention to students from low social economic backgrounds.
- Schools should set up student support systems and focus much attention on students from low social economic backgrounds.
- Schools should identify needy students and assist them with financial aid or grant them scholarship status.
- Public educational institutions (first cycle) should set up functioning counseling centers in all public basic schools and ensure that, adequate counselors are made available to the students.

- Students from unstable or broken families should seek professional advice from counselors on how to manage their psychological problems and counselors should provide the necessary assistance to them.

**REFERENCES**


Aman, K., and Suruchi, S. (2011). A Comparative Study of Classification Algorithms for Spam Email Data Analysis", International Journal on Computer Science and Engineering, May 2011 ,Vol. 3 No. 5, pp 1890-1895

Anamuah-Mensah, J. (1997). Native science beliefs among some Ghanaian students. International Journal of Science Education, 20 (1), 115-124.

Anamuah-Mensah, J. (1999). *Science and Technology Education in Ghana.* A paper delivered at the National education Forum on the theme: Towards Sustaining an Effective National Education System, held at the Accra International Conference Centre, Accra, 17-19th November.

Anderson, G., & Benjamin, D. (1994). The determinants of success in university introductory economics courses. Journal of Economic Education, 25(2), 99 – 119

Aughinbaugh, A., C.R. Pierret, and D.S. Rothstein (2005). "The Impact of Family Structure Transitions on Youth Achievement: Evidence from the Children of the NLSY79." Demography, 42 (3): 447-468.

Allison, P.D. and Furstenberg, F.F (1989). "How Martial Dissolution Affects Children: Variation by Age and Sex." Developmental Psychology, 25, 540-549.

Al-Radaideh, Q., Al-Shawakfa, E. and Al-Najjar, M. (2006) 'Mining Student Data Using Decision Trees', The 2006 International Arab Conference on Information Technology (ACIT'2006).

Akyeampong K (2005) "Vocationalised Secondary Education in Ghana" in Lauglo J & MacLean R (Eds) Vocationalisation of Secondary Education Revisited, Springer, Dordrecht, The Netherlands

Akyeampong, K. (2008) Educational Expansion and Access in Ghana: A Review of 50 Years of Challenge and Progress. Available at, http://www.create-rpc.org/pdf_documents/50%20 Years%20of %20Educational_Progress_%20in_Ghana.pdf. Accessed on November 1st, 2014.

Acheampong, et al (2012). Access, Transitions and Equity in Education in Ghana: Researching Practice, Problems and Policy. Creative pathway to access, Research Monograph No 72. Available at http://www.create-rpc.org/pdf_documents/PTA72.pdf. Accessed on Accessed on November 1st, 2014.

Addae-Mensah, I., Djangmah, J. & Agbenyega, C (1973) Family Background and Educational opportunity in Ghana: Accra: Ghana Universities Press

Addae-Mensah, I (2000) Education in Ghana: A Tool for Social Mobility or Social Stratification. The J.B. Danquah Memorial Lectures, April 2000.

Ayesha, S. , Mustafa, T. , Sattar, A. and Khan, I. (2010) 'Data Mining Model for Higher Education System', European Journal of Scientific Research, vol. 43, no. 1, pp. 24-29.

Armstrong, J. and Anthes, K. (2001). How data can help. American School Board Journal 188(11), 38-41.

Alwin, L. (2002). The will and the way of data use: School Administrator. 59(11), 11.

Aripin, R., Mahmood, Z., Rohaizad, R., Yeop, U., &Anuar, M. (2008). Students' Learning Styles And Academic Performance. 22nd Annual SAS Malaysia Forum, Kuala Lumpur Convention Center.

Available. at, http://www.sas.com/offices/asiapacific/malaysia/academic/2008_sum_ paper/Rasimah Zurina.pdf. Accessed on October 8th, 2013.

American Association of School Administrators (2002). Using data to improve schools: What's working. Available at; www.aasa.org/cas/UsingDataToImproveSchools.pdf. Accessed on March 14th, 2014.

Bisnaire .L, Firestone p. & Rynard D. (1990). "Factors Associated with Academic Achievement in Children Following Parental Separation."American Journal of Orthopsychiatry, 60(1),

Bond, C.R., and McMahon, R. J, (1984). "Relationships between Marital Distress and Child Behavior Problems, Maternal Personal Adjustment, Maternal Personality and Maternal Parenting Behavior", Journal of Abnormal Psychology, 93, 348-351.

Battle, Juan. and Michael Lewis. 2002. The increasing significance of class: The relative effects of race and socioeconomic status on academic achievement. *Journal of Poverty*, *6*(2), 21-35.

Bachman, Sarah L. 2000. "The Political Economy of Child Labor and Its Impacts on International Business." Business Economics, vol. 35, no. 3 (July), pp. 30-41.

Brown, S.L. (2004). "Family Structure and Child Well-Being: The Significance of Parental Cohabitation." Journal of Marriage and Family, 66, 351-367.

Baker, R., Corbett, A., Koedinger, K. (2004). Detecting student misuse of intelligent tutoring systems. In International Conference on Intelligent Tutoring Systems, Alagoas, Brazil, 531–540.

Baker, R., Yacef, K. (2009) The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1, 1, 3-17.

Baker, R (2010). Data Mining for Education to appear in McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier.

Baradwaj, B. and Pal, S. (2011) 'Mining Educational Data to Analyze Student s' Performance', International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63-69.

Burns, N. and Grove, S.K. (1993) The practice of nursing research conduct, critique & utilization (second edition), Philadelphia: W.B. Saunders.

Campbell, J., &Oblinger, D. (2007).Academic analytics. Washington, DC: Educause.

Cochran, W. G., (1953). Sampling Techniques. Second Edition. John Wiley & Sons, Inc. New York. Library Of Congress Catalog Card Number: 63-7553

Calders, T., & Pechenizkiy, M. (2012).Introduction to the special section on educational data mining.SIGKDD Explor.Newsl.,13(2), 3-6.

Corbett, A. T. (2001). Cognitive Computer Tutors: Solving the Two-Sigma Problem. Paper presented at the Proceedings of the 8th International Conference on User Modeling 2001.

Crosnoe, R., Monica, K. J and Glen, H .E .Jr. (2004). School size and the interpersonal side of education: An example of Race/Ethnicity and organizational context. Social Science Quarterly, 85(5).

Castro, F., Vellido, A., Nebot, A. Mugica, F. (2007).Applying Data Mining Techniques to e-Learning Problems.

Chrispeels, Brown, and Castillo (2000). School leadership teams: Factors that influence their development and effectiveness. Advances in Research and Theories of School Management and Educational Policy, 4, 39-73.

Chrispeels, J. H. (1992). Purposeful restructuring: Creating a climate of learning and achievement in elementary schools. London: Falmer Press.

Considine, G. &Zappala, G. (2002). Influence of social and economic disadvantage in the academic performance of school students in Australia. Journal of Sociology, 38, 129-148.

Crommey, A. (2000). Using Student Assessment Data: What Can We Learn From Schools? Oak Brook, IL: North Central Regional Educational Laboratory.

Chandra, E. and Nandhini, K. (2010) 'Knowledge Mining from Student Data', European Journal of Scientific Research, vol. 47, no. 1, pp. 156-163.

Donge J (2003) Into the black box of Ghanaian Education: why do increased inputs not lead to better educational outputs? Mimeo, The Hague: ISS

Deci, E.L. (1981). An Instrument to Assess Adults' Orientations Towards Control Versus Autonomy with Children: Relation on Intrinsic Motivation and Perceived Competence. Journal Of Educational Psychology, 73,642-650

Devadoss, S., & Foltz, J. (1996). Evaluation of factors influencing students attendance and performance. American Journal of Agricultural Economics, 78(3), 499 – 507

Doyle, D. P. (2003). Data-driven decision making: Is it the mantra of the month or does it have staying power? T.H.E. Journal, 30(10), 19-21.

David, H. and Heikki M.P. (2001) Principles of Data Mining**.** The MIT Press, Cambridge, MA. (page 124)

Diamond, J.B. & Spillane, J.P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? Teachers College Record, 106(6), 1145-1176.

Dunham, M. (2003).Data Mining: Introductory and Advanced Topics. Upper Saddle River, NJ: Pearson Education.

El-Halees, A. (2008) 'Mining Students Data to Analyze Learning Behavior: A Case Study', The 2008 international Arab Conference of Information Technology (ACIT2008). Available at, https://uqu.edu.sa/files2/tiny_mce/plugins/filemanager/files/30/papers/f158.pdf. Accessed on September 17th, 2013.

Elkan, C. (2012). Evaluating Classifiers. Available at, http://cseweb.ucsd.edu/~elkan/250 Bwinter2012/classifiereval.pdf. Accessed on August 10th, 2014.

Eamon, M.K (2005). Social demographic, school, neighborhood, and parenting influences on academic achievement of Latino young adolescents. Journal of Youth and Adolescence, 34(2), 163-175.

Escarce, J. J (2003).Socioeconomic status and the fates of adolescents. Available at, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid. Accessed on December 12th, 2013.

Earl, L., & Katz, S. (2002). Leading schools in a data-rich world. In K. Leithwood & P. Hallinger (Eds.), Second international handbook of educational leadership and administration (pp. 1003–1024).

Earl, L., & Katz, S. (2006). Leading schools in a data-rich world. Thousand Oaks, CA: Corwin Press.

El-Halees, A. (2009). "Mining students data to analyze e-Learning behavior: A Case Study". Available at, https://uqu.edu.sa/files2/tiny_mce/plugins/filemanager/files/30/papers/f158.pdf. Accessed on August 30th, 2014.

Feldman, J. & Tung, R. (2001).Using data-based inquiry and decision making to improve instruction. ERS Spectrum 19(3), 10-19.

Fraenkel, J.R., &Wallen, N.E. (1996). How to design and evaluate research in education (3rd ed.). New York: McGraw-Hill.

Fawcet, T. (2003). ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, Intelligent Enterprise Technologies Laboratory HP Laboratories Palo Alto. Available at, http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf. Accessed on August 16th, 2014.

Fomby, Paula, and Andrew J. Cherlin. 2007. "Family Instability and Child Well-Being." American Sociological Review 72:181-204.

Gershoff, E. T., Aber, J. L., Raver, C. C., & Lennon, M. C. (2007). Income is not enough: Incorporating material hardship into models of income associations with parent mediators and child outcomes. Child Development, 78(1), 70-95.

Gottman, J.M. & Katz, L.F. (1989). "The Effects of Marital Discord on Young Children's Peer Interaction and Health", Developmental Psychology, 25,373-381.

Government of Ghana (GOG, 2004) White Paper on the Report of the Education Reform Review Committee, Ministry of Education Youth and Sports

Government of Ghana (GOG) (2002). *Meeting the Challenges of Education in the Twenty First Century: Report of the President's Committee on Review of Education Reforms in Ghana.* Accra: Adwinsa Publications Ltd.

Guidubaldi, J. & Perry, J.D. (1985). "Divorce and Mental Health Sequelae For Children: A Two Year Follow-Up of A Nationwide Sample," Journal of the American Academy Of Child Psychiatry, 24, 531-537.

Gardner, M.J. and Altman, D.G. (1989). Calculating confidence intervals for proportions and their differences. London: BMJ Publishing Group.

Graetz, B. (1995), Socio-economic status in education research and policy in John Ainley et al., Socio-economic Status and School Education DEET/ACER Canberra.

Guan, J., Nunez, W., & Welsh, J. (2002). Institutional strategy and information support: the role of data warehousing in higher education. Campus-Wide Information Systems, 79(5), 168174.

Hetherington, E.M, Cox, M., & Cox, R, (1979). “Family Interaction and Social Emotional and Cognitive Development of Children Following Divorce".

Heyman, S. P. (1980). Students Learning in Uganda.Textbook available and other factors.Comparative Education Review, 24.2.

Hand, D., Mannila, H., and Smyth, P. (2001) Principles of Data Mining. The MIT Press. P. ISBN: 026208290x

Hijazi, S.T. and Naqvi, S.M.M. Raza. (2006). ‘Factors Affecting Students’ Performance: A Case of Private Colleges’. Bangladesh e-Journal of Sociology: Volume 3, Number 1.

Hassan, T. (1983) Psychosocial predictors of academic achievement. Psychology for Everyday Living, 2 (2), 155 - 169.

Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd edition). Morgan Kaufmann Publishers.

Hansen, N.M and Mastekaasa, A. (2006). Social origins and academic performance at university. Oxford University press. Available at http://esr.oxfordjournals.org/cgi/content/abstract/22/3/277. Accessed on November 17th, 2013

Ingram, D. Louis, K.S., Schroeder, R.G. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. Teachers College Record, 106(6), 1258-1287.

‘Information & entropy’ Available at, http://www.csun.edu/~twang /595DM/Slides/ Information %20&%20Entropy.pdf. Accessed on July 19th, 2014.

Johnson, J. H. (1999). Educators as researchers. Schools in the Middle, 9(1), 38-41.

Johnson, J. H. (2000). Data-driven school improvement. Journal of School Improvement, 1(1). Available at, http://www.ncacasi.org/jsi/2000v1i1/data_driven. accessed on January 13th, 2014.

Jeynes, W. H. (2002). Examining the effects of parental absence on the academic achievement of adolescents: The challenge of controlling for family income.Journal of Family and Economic Issues, 23(2), 56-65.

Kwesiga, C.J. (2002). Women’s access to higher education in Africa: Uganda’s experience. Kampala: Fountain publishers Ltd.

Kennedy, E. (2003). Raising Test Scores for All Students: An Administrator’s Guide to Improving Standardized Test Performance. Thousand Oaks, CA: Corwin Press.

Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., &Haydock, M. (2012). Analytics: The Widening Divide. MIT Sloan Management Review, 53(2), 1-22.

Kotsiantis, S., Pierrakeas, C. and Pintelas, P. (2003). Preventing student dropout in distance learning systems using machine learning techniques.

Lachat, M.A. (2002). Data-Driven High School Reform: The Breaking Ranks Model. Providence, RI: Northeast and Islands Regional Educational Laboratory at Brown University.

Lafee, S. (2002). Data-driven districts. School Administrator, 59(11), 6-7, 9-10, 12, 14-15.

Luan, J. (2000). "An Exploratory Approach to Data Mining in Higher Education: A Primer and a Case Study." Paper presented at the AIR Forum, Seattle, Wash.

Luan, J. (2002). Data Mining and Knowledge Management in Higher Education – Potential Applications. Paper presented at the Annual Forum for the Association for Institutional Research, Toronto, Ontario, Canada. Available at, http://eric.ed.gov/ERICWebPortal/detail?_accno=ED474143. Accessed on August 20th, 2013.

Larose, T.D. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. ISBN 0-471-66657-2 Copyright C _ 2005 John Wiley & Sons, Inc.

Likert, R. (1932). .A technique for the measurement of attitudes..Archives of Psychology, No. 140

Leventhal, B. (2010). An introduction to data mining and other techniques for advanced analytics. Journal of Direct, Data and Digital Marketing Practice, 12(2), 137-153.

Maimon, O. and Rokach, L. (2005) Data Mining and Knowledge Discovery Handbook (2nd Edition). ISBN 978-0-387-09822-7 Library of Congress Control Number: 2010931143.

Myers, R. and Walpole, R. (1978). Tests of Hypotheses: Probability and Statistics for Engineers and Scientists, (2nd Edition), Macmillan Publishing Co., Inc., New York, NY, 1978, pp. 268 - 273.

Marks, G. (2006). Family size, family type and student achievement: Cross-national differences and the role of socioeconomic and school factors. Journal of Comparative Family Studies, 37(1), 1-27.

Maani (1990).Factors Affecting Academic Performance.Unpublished dissertation, Makerere University, Kampala, Uganda.

Mulkey, L., Crain, R., & Harrington, A. (1992). One-parent households and achievement: Economic and behavioral explanations of a small effect. Sociology of Education, 65,48-65.

Mouton, J. 1996.Understanding social research. Pretoria: Van Schaik.

Minaei-Bidgoli, B., Kashy, D.A., Kortmeyer, G. and Punch, W.F. (2003). Predicting student performance: an application of data mining methods with an educational Web-based system, 33rd Annual Frontiers in Education, vol. 1, pp.T2A–18

McMillan, J and Western ,J. (2000). Measurement of Social-Economic Status of Australian Higher Education Students. Higher Education, Vol.39, No. 2. Springer.

Montgomery, D.C., Peck, E.A., and Vining, G.G. (2001) Introduction to linear regression analysis. John Wiley & Sons, Inc., New York NY.

McIntire, T. (2002).The administrator's guide to data-driven decision making, Technology and Learning, 22(11), 18-33.

Mason, S. (2002). Turning data into knowledge: Lessons from six Milwaukee public schools.

Nemati, H., & Barko, C. (2004). Organizational Data Mining (ODM): An Introduction. In H. Nemati& C. Barko (Eds.), Organizational Data Mining (pp. 1-8). London: Idea Group Publishing.

Neville, P. (1999), "Growing Trees for Stratified Modeling," Computing Science and Statistics, 30.

Nsiah-Gyabaah, K. (2009) The Missing Ingredients in Technical and Vocational Education in Meeting the Needs of Society and Promoting Socio-Economic Development in Ghana; Journal of Polytechnics in Ghana; Volume 3, No. 3

Osunloye, A. (2008) Family Background and Student Academic Performance. Available at, http://socybertycom/education/family-background-and-student-academic-performance. Accessed on December 15th, 2013.

Pardini, P. (2000). Data, well done. Journal of Staff Development 21(1), 12-18.

Protheroe, N. (2001). Improving teaching and learning with data-based decisions: Asking the right questions and acting on the answers. ERS Spectrum 19(3), 4-9.

Petrides, L., & Nodine, T. (2005).Anatomy of school system improvement: Performance-driven practices in urban school districts. San Francisco, CA: Institute for the Study of Knowledge Management in Education and NewSchools Venture Fund.

Pedrosa, et al (2006). Educational and social economic background of undergraduates and academic performance: consequences for affirmative action programs at a Brazilian research university. Available at, http://www.comvest.unicamp.br/paals/artigo2.pdf. Accessed on Feb 15th, 2014.

Peter, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., and Rudiger, W. (2000) CRISP–DM Step-by-Step Data Mining Guide. Available at, http://www.crisp-dm.org/. Accessed on January 20th, 2014.

Quinlan, J.R. (1986). Induction of Decision Tree, Journal of Machine learning, Morgan Kaufmann Vol.1, 1986, pp.81-106.

Quinlan, J.R. (1992). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc.

Quinlan, J.R. (1996). Learning decision tree classifiers. ACM Computing Surveys (CSUR), 28(1):71–72.

Romer, D. (1993). "Do students go to class? Should they?", Journal of Economic Perspectives, Summer, 7(3), 167-174.

Romero, C. and Ventura, S. (2007) 'Educational data Mining: A Survey from 1995 to 2005', Expert Systems with Applications (33), pp. 135-146.

Romero, C., & Ventura, S. (2010). Educational Data Mining: The Journal of Information and Knowledge Management Systems, 37(4), 502-515.

Polit, DF & Hungler, BP 1993: Essentials of nursing research:Methods, appraisal, and utilization; 3rd edition. Philadelphia:Lippincott.

Ranjan, J., & Malik, K. (2007). Effective educational process: a data-mining approach. VINE: The Journal of Information and Knowledge Management Systems, 37(4), 502-515.

Raychaudhuri, Amitava, Debnath, Manojit, Sen, Seswata&Majundra, brajaGopal. (2010). Factors affecting Student's academic performance: A case study in Agartala municipal concial area. Bangladesh e-journal of sociology, vol.7, Number.2

Rutter, M. (1978). Early Sources of Security and Competence. In J.S Bruner & A Garten, (eds.), Human Growth and Development (pp.323-61) London: Oxford University Press.

Stringfield, S., Wayman, J. C., &Yakimowski, M. (2003, March).Scaling up data use in classrooms, schools, and districts. Paper presented at the Scaling Up Success conference, Harvard Graduate School of Education, Boston MA.

Sentamu, N.P. (2003). School's influence of learning: A case of upper primary schools in Kampala &Wakiso Districts.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75(3), 417 – 453. Available at, http://steinhardt.nyu.edu/scmsAdmin/media/users/lec321/Sirin_Articles/Sirin_2005.pdf. Accessed on August 30th, 2014.

Sogbetun, A. A. (1981) Teachers and students opinion about the causes of poor academic performance in secondary schools Unpublished M.Ed. Project. University of Ibadan.

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of communication. Urbana, IL: University of Illinois Press.

Streifer, P.A. (2002, July). Data-driven decision-making: What is knowable for school improvement.

Statistics, Research, Information Management and Public Relations (SRIMPR) 2012, Basic Statistics and Planning Parameters for Basic Education in Ghana. Available at, http:// www.moe.gov.gh/moe/docs/Basic%20Report%202011-2012%20No.1.pdf. Accessed on October 9th 2013.

Sparkes, J. (1999), Schools, Education and Social Exclusion, CASE Paper 29, Centre for Analysis of Social Exclusion, London School of Economics, London.

Santor, D.A., Deanna, M. and Vivek, K. (2000). Measuring peer pressure, popularity, and conformity in adolescent boys and girls: Predicting school performance, sexual attitudes, and substance abuse. Journal of Youth and Adolescence 29(2) 163.

Supovitz, J., & Taylor B.S. (2003).The Impact of Standards-based Reform in Duval County, Florida, 1999-2002. Philadelphia, PA: Consortium for Policy Research in Education.

Shannaq, B. , Rafael, Y. and Alexandro, V. (2010) 'Student Relationship in Higher Education Using Data Mining Techniques', Global Journal of Computer Science and Technology, vol. 10, no. 11, pp. 54-59.

Slavin, R.E. (2002). Evidence-based education policies: transforming educational practice and research. Educational researcher, 31(7), 15-21.

Sarantakos, S. (1997). Social research. New York: Palgrave Publishers Ltd.

Togneri, W., & Anderson, S. (2003). Beyond Islands of Excellence: What Districts Can Do to Improve Instruction and Achievement in All Schools. Washington, DC: Learning First Alliance

Thorn, C. A. (2001). Knowledge management for educational information systems: What is the state of the field? Education Policy Analysis Archives 9(47), 1-32. Available at: http://epaa.asu.edu/epaa/v9n47/. Accessed on March 5th, 2014.

Ushie, M. A., Ononogha, G. I., Owolabi, E. O. &Emeka, J. O. (2012) Influence of Family Structure on Students' Academic Performance in Agege Local Government Area, Lagos State, Nigeria.

Vialardi, C., Chue, J., Peche, J. P., Alvarado, G., Vinatea, B., Estrella, J., & Ortigosa, A. (2011). A data mining approach to guide students through the enrollment process based on academic performance. User Modeling and User - Adapted Interaction, 27(1-2), 217-248.

Vogt, W. Paul (1999).Dictionary of statistics and methodology. Sage: Thousand Oaks, California.

Wang, F.-H. (2008). Content Recommendation Based on Education-Contextualized Browsing Events for Web-Based Personalized Learning, Educational Technology & Society, 11(4), 94-112

Wang, Y.-h., & Liao, H.-C.(2011). Data mining for adaptive learning in a TESL-based e- learning system.Expert Systems with Applications, 38(6), 6480-6485.

Wallerstein (1980) Surviving the Break up: How Children and Parents Cope with Divorce. New York: Basic Books.

Wallerstein, J.S., Corbin, S.B., & Lewis, J.M. (1988) "Children of Divorce: A Ten-Year Study.

Weissman (1983). "The Depressed Mother and her Rebellious Adolescents. In H. Morrison (Ed.) Children of Depressed Parents Risk, Identification, and Intervention (pp. 99-113). New York: Grun & Stratton.

Wohlstetter, P., Van Kirk, A. N., Robertson, P. J., & Mohrman, S. A. (1997). Organizing for successful school-based management. Alexandria, VA: Association for Supervision and Curriculum Development.

Witten, I.H. and Eibe, F. (2005). Data Mining: Practical Machine Learning Tools and Techniques, (2nd Edition). Morgan Kaufmann Publishers, 500 Sansome Street, Suite 400, San Francisco, CA 94111.

Yates, D.S., David, S.M. and Daren, S.S. (2008). The Practice of Statistics, (3rd Edition) Freeman. ISBN 978-0-7167-7309-2.

Yamane. T. (1967). Elementary Sampling Theory. Prentice-Hall, Inc., Englewood Cliffs, N.J. Raj, D. (1972). The Design Of Sample Surveys. McGraw-Hill Book Company, New York.

Zoubin, G. (2000). Encyclopedia of cognitive science: information, entropy, communication, coding, bit, learning. ©Macmillan Reference Ltd. University College London United Kingdom

Zou, Y., An, A., Huang, X. (2005) „Evaluation and automatic selection of methods for handling missing data", IEEE International Conference on Granular Computing, vol.2, pp. 728- 733.

## APPENDIX A

**Socio-economic status and academic performance of students in public basic schools in the Ablekuma West constituency of the Greater Accra region.**

**Students Questionnaire**

As part of the requirements for the fulfillment of the requirements for master's degree in Information Technology (I.T), I am conducting a research in the development of a data mining classification model to predict the academic performance of students in public basic schools in the Ablekuma west constituency using socio-economic variables. I am therefore requesting for your participation in this study by filling this questionnaire. All answers provided will be strictly confidential. The information provided will be beneficial for addressing the challenges of students from poor socio-economic homes.

**Instructions for section A**

*Please fill in as appropriate*

**SECTION A: STUDENT DEMOGRAPHIC DATA**

1. **Age** ☐
2. **Gender**
   a. Male ☐ b. Female ☐
3. **Class**
   a. JHS 1 ☐
   b. JHS 2 ☐

**Instructions for section B**

*Please tick as appropriate*

| | | Strongly disagree | Agree | Strongly agree | Strongly agree |
|---|---|---|---|---|---|
| | **Question** | **4** | **3** | **2** | **1** |
| **1** | I come from a large family background | | | | |
| **2** | My parents are educated | | | | |
| **3** | My parents are gainfully employed | | | | |
| **4** | My parents are married | | | | |
| **5** | My parents are high income earners | | | | |
| **6** | My academic performance can be rated as good | | | | |