

**DSpace Institution**

**DSpace Repository**

<http://dspace.org>

---

Information Technology

thesis

---

2020

# PREDICTING MATERNAL MORTALITY IN AMHARA REGION USING DATA MINING TECHNIQUES

Kindie, Abinet

---

<http://hdl.handle.net/123456789/11278>

*Downloaded from DSpace Repository, DSpace Institution's institutional repository*



**BAHIR DAR UNIVERSITY**  
**BAHIR DAR INSTITUTE OF TECHNOLOGY**  
**SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES**  
**COMPUTING FACULTY**

**PREDICTING MATERNAL MORTALITY IN AMHARA REGION**  
**USING DATA MINING TECHNIQUES**

**By**  
**Abinet Kindie**

**July 2020/**  
**Bahir Dar, Ethiopia**

PREDICTING MATERNAL MORTALITY IN AMHARA REGION USING DATA MINING  
TECHNIQUES

By

Abinet Kindie

“A thesis submitted to Bahir Dar Institute of Technology in partial fulfilment of the requirements for the degree of Masters of Science in Information Technology in the computing faculty.”


Advisor: Abebe Tesfahun (PhD)

July 2020/ Bahir Dar, Ethiopia

### Declaration

I declare that this thesis is my original work and that has not been presented for a degree from any other university and I have acknowledged all the materials used in this work.

Name of the student: Abinet Kindie

Signature: 

Date of submission: 6/24/2020

Place: Bahir Dar

This thesis has been submitted for examination with my approval as a university advisor.

Advisor Name: Abebe Tesfahun (PhD)

Advisor's Signature:





©2020

**Abinet Kindie Wondim**  
**ALL RIGHTS RESERVED**

## Dedication

This research work is dedicated to my beloved family for their encouragement and support.

## **Acknowledgement**

Foremost, I would like to thank God he made all things possible. It is a great pleasure for me to express my heartfelt gratitude to my advisor Dr. Abebe Tesfahun who gave me constructive ideas on my research. Thank you so much for guiding me in a proper direction to achieve the objectives of this research.

I would like to express heartfelt thanks to Mr. Birhanu Haile and Mr. Alemu Kumilachew, for their constructive comments for this research work.

My Grateful thanks to the Debre Markos referral hospital, Felege Hiwot referral hospital, and Addis Alem Hospital staff members who help me during data collection process.

I would like to extend my thanks to Bahir Dar University for sponsoring this thesis research.

My special thanks go to my beloved family for their support and encouragement. Especially, to my sister Yeshiharge Kindie who has always encouraging me for higher success. “Thank you”.

There are many individuals that have really contributed directly or indirectly for the successful accomplishment of this research, and all of them deserve special appreciation and acknowledgement for being with me in all those challenging times of the study.



## **Abstract**

Maternal mortality is the death of pregnant women due to complications arising during the period of pregnancy and after delivery. Predicting maternal mortality and identifying the major determinants for maternal mortality are important for decision making during treatment and follow-up. Therefore, this research is aimed to apply data mining techniques to build a model that can assist in predicting maternal mortality. The data is taken from Debre Markos Referral Hospital, Felege Hiwot Specialized Referral Hospital, and Addis Alem Hospital located in Amhara regional state. The six-step hybrid knowledge discovery method is employed as a framework for the activities done in this study. Dataset pre-processing was applied for missing value handling, noise removal, and data transformation. In this study, RStudio data mining tool and classification-based data mining techniques such as decision tree, naïve Bayes, and Support vector machine were employed to build the predictive model. The performances of the models were evaluated using sensitivity, specificity, precision, recall, and Accuracy. 10-fold cross-validation and percentage split were adopted as the test option to check the performance of each classifier. Experimental results show that the most effective model to predict the status of maternal outcome and determinant factors of maternal death appears to be the Pruned J48 decision tree model with a classification accuracy of 97.56%, precision 96.83%, sensitivity 99.29%, specificity 94.78%, and recall 99.29%. The extracted rules from the selected models show that the maternal condition, Age, and obstetric complications were the major determinant factors of maternal mortality. Postpartum Haemorrhage (PPH), Eclampsia, and Antepartum Haemorrhage APH) were the major complications that have a high impact on maternal mortality. The outcome of the study can be used as an assistant tool for physicians to make a more consistent diagnosis of factors that causes maternal mortality. The possibility of integrating the results of this study with a knowledge-based system should be discovered so that domain experts can access the system in their problem solving and decision-making tasks.

**Key words:** maternal mortality, data mining, predictive modeling, determinant factor

## **Table of Contents**

<b>Acknowledgement</b> .....	<b>vi</b>
<b>Abstract</b> .....	<b>vii</b>
<b>Table of Contents</b> .....	<b>viii</b>
<b>List of Abbreviations</b> .....	<b>xi</b>
<b>List of Figures</b> .....	<b>xiii</b>
<b>List of Tables</b> .....	<b>xiv</b>
<b>1. Introduction</b> .....	<b>1</b>
<b>1.1. Background of the Study</b> .....	<b>1</b>
<b>1.2. Statement of The Problem</b> .....	<b>2</b>
<b>1.3. Objectives of the Study</b> .....	<b>3</b>
<b>1.3.1. General Objective</b> .....	<b>3</b>
<b>1.3.2. Specific Objective</b> .....	<b>4</b>
<b>1.4. Scope and Limitation of the Study</b> .....	<b>4</b>
<b>1.5. Significance of the Study</b> .....	<b>4</b>
<b>1.6. Methodology of the Study</b> .....	<b>5</b>
<b>1.6.1. Research Design</b> .....	<b>5</b>
<b>1.6.2. Literature Review</b> .....	<b>6</b>
<b>1.6.3. Data Collection Technique</b> .....	<b>6</b>
<b>1.6.4. Implementation Tool</b> .....	<b>6</b>
<b>1.7. Organization of the Thesis</b> .....	<b>6</b>
<b>2. Literature Review</b> .....	<b>8</b>
<b>2.1. Overview of Data Mining</b> .....	<b>8</b>
<b>2.1.1. Data Mining</b> .....	<b>8</b>
<b>2.2. Data Mining Process Models</b> .....	<b>9</b>
<b>2.2.1. Knowledge Discovery in Database (KDD)</b> .....	<b>10</b>
<b>2.2.2. The CRISP-DM Process</b> .....	<b>11</b>
<b>2.2.3. SEMMA Process Model</b> .....	<b>13</b>
<b>2.2.4. Hybrid Model</b> .....	<b>14</b>

<b>2.3.</b>	<b>Data Mining Models and Tasks .....</b>	<b>17</b>
2.3.1.	<b>Classification .....</b>	<b>17</b>
<b>2.4.</b>	<b>Application of Data Mining Techniques .....</b>	<b>18</b>
2.4.1.	<b>Application of Data Mining in Healthcare .....</b>	<b>19</b>
<b>2.5.</b>	<b>Related works .....</b>	<b>20</b>
<b>3.</b>	<b>Methodology of the Study .....</b>	<b>24</b>
3.1.	<b>Research Design.....</b>	<b>24</b>
3.2.	<b>Tools.....</b>	<b>26</b>
3.3.	<b>Algorithms for Model Building.....</b>	<b>27</b>
3.3.1.	<b>Decision Tree Classifier .....</b>	<b>28</b>
3.3.1.1.	<b>Decision tree building.....</b>	<b>28</b>
3.3.1.2.	<b>Pruning decision tree .....</b>	<b>30</b>
3.3.1.3.	<b>Rule induction.....</b>	<b>30</b>
3.3.2.	<b>Naïve Bayes Classifier.....</b>	<b>31</b>
3.3.2.1.	<b>Naïve Bayes Theorem.....</b>	<b>32</b>
3.3.3.	<b>Support Vector Machine Classifier .....</b>	<b>32</b>
3.4.	<b>Model Evaluation Techniques.....</b>	<b>33</b>
3.4.1.	<b>Methods of Training and Testing the Model.....</b>	<b>33</b>
3.4.2.	<b>Performance Evaluation of Classification Models.....</b>	<b>34</b>
3.4.2.1.	<b>Confusion Matrix .....</b>	<b>34</b>
<b>4.</b>	<b>Data Understanding and Pre processing .....</b>	<b>36</b>
4.1.	<b>Data Understanding .....</b>	<b>36</b>
4.1.1.	<b>Data Collection Method.....</b>	<b>38</b>
4.2.	<b>Data Preprocessing.....</b>	<b>38</b>
4.2.1.	<b>Data Cleaning.....</b>	<b>39</b>
4.2.2.	<b>Data Transformation .....</b>	<b>41</b>
4.2.2.1.	<b>Data Discretization.....</b>	<b>41</b>
4.2.3.	<b>Selecting the Attributes .....</b>	<b>42</b>
4.2.4.	<b>Data Formatting.....</b>	<b>44</b>
4.3.	<b>Experimentation and Discussion .....</b>	<b>46</b>
4.3.1.	<b>Model Building.....</b>	<b>46</b>

4.3.1.1. Model Building Using J48 Algorithm.....	48
4.3.1.2. Model Building Using Naïve Bayes Algorithm.....	51
4.3.1.3. Model Building Using Support Vector Machine Algorithm .....	52
4.3.2. Model Comparison.....	54
4.3.3. Rule Extraction .....	56
5. Conclusion and Recommendation.....	62
5.1. Conclusion .....	62
5.2. Recommendation.....	64
REFERENCES.....	65
APPENDEXES.....	69
Appendix 1.....	69
Appendix 2.....	70
Appendix 3.....	71
Appendix 4.....	72
Appendix 5.....	74
Appendix 6.....	75
Appendix 7.....	77

## List of Abbreviations

Acc-----	Accuracy
APH-----	Antepartum Haemorrhage
ARFF-----	Attribute Relation File Format
CRISP-DM-----	Cross Industry Standard Process for Data Mining
CS -----	Caesarean section
CSV-----	Comma Separated Value
CV-----	cross validation
DM-----	Data mining
EDHS-----	Ethiopian Demographic and Health Survey
FMOH-----	Federal ministry of health
FN-----	False negative
FP -----	False positive
htr-----	HIV test Result
I-miner-----	intelligent miner
KDP-----	Knowledge discovery process
mcondition-----	maternal condition
MDSR-----	Maternal death surveillance and Response
ML-----	Machine learning
MLP-----	Multi-layer perceptron
MM-----	Maternal mortality
mod-----	Mode of delivery
mstatus-----	Maternal status
NA-----	Not available
nb -----	naïve Bayes
NR -----	Non-Reactive

OBS-----obstetric  
PPH -----Postpartum Haemorrhage  
R -----Reactive  
SDGs -----Sustainable Development Goals  
SEMMA -----Sample, Explore, Modify, Model, Assess  
SMOTE -----Synthetic Minority over Sampling Technique  
Spl-----split  
SVD -----Spontaneous Vaginal Delivery  
SVM-----Support Vector Machine  
TN-----True negative  
TP-----True positive  
UNICEF-----United Nations Children Fund  
Weka-----Waikato Environment for Knowledge Analysis  
WHO----- World Health Organization

## List of Figures

Figure 2.1: Data Mining Architecture with Other interdisciplinary Fields (Han and Kamber 2006)	9
Figure 2.2: The KDD process (Fayyad et al.1996)	11
Figure 2.3: The general architecture of CRISP-DM (Wirth, n.d.)	12
Figure 2.4: SEMMA Process Model (Wirth, n.d.)	14
Figure 2.5: The Six steps of Hybrid KDP Model (Cios and Kurgan n.d.)	16
Figure 2.6: Data mining Models and tasks (Fayyad et al., 1996)	18
Figure 2.7: Process of data mining in healthcare (Thorat & Kute, 2014)	20
Figure 3.1: Graphical Overview of Overall Research Design and Methodology	25
Figure 3.2: Architecture of predictive model for maternal mortality	26
Figure 3.3: RStudio user interface	27
Figure 3.4: A simple decision tree for two class classification (Han and Kamber 2006)	30
Figure 4.1: Data pre-processing steps for quality data mining	39
Figure 4.2: CSV files used for maternal mortality prediction	45
Figure 4.3: Side by side view of the class variable using SMOTE	47
Figure 4.4: Pruned J48 classifier detailed in Rweka interface	51
Figure 4.5: Naive Bayes classifier detailed	52
Figure 4.6: SVM classifier detailed	54
Figure 4.7 : visualization of performance comparison the three models	55
Figure 4.8: Performance measures of the selected model	56
Figure 4.9: Predictive relationship between maternal status and maternal condition	59
Figure 4.10: Predictive relationship between maternal status and obstetric complication	60
Figure 4.11: Predictive relationship between maternal status and maternal age	61

## List of Tables

Table 3.1:Confusion matrix of two class classification result .....	34
Table 4.1: list of Attributes in the initial dataset and its descriptions.....	37
Table 4.2: list of attributes with missing and estimated values .....	40
Table 4.3: Data discretization .....	42
Table 4.4: List of selected attributes and their descriptions.....	43
Table 4.5: List of selected attributes based on their gain ratio value.....	44
Table 4.6 : Experiments and schemes.....	47
Table 4.7: different parameters and its value for J48 decision model building .....	49
Table 4.8: Experimental results of J48 decision tree classifier.....	50
Table 4.9: Experimental results of Naive Bayes classifier .....	52
Table 4.10: Experimental results of SVM classifier .....	53
Table 4.11 : Performance summary of the three Models.....	55
Table 4.12 : Confusion Matrix result of the selected J48 -C 0.5 -M 2 .....	56



# CHAPTER ONE

## 1. Introduction

### 1.1. Background of the Study

Identifying the leading causes of death of the mother is the most serious issue in the planning of health care interventions (Mehta and Bhatt 2016). Antenatal and postnatal cares are the most effective health interventions to prevent maternal morbidity and mortality in the place where the general health status of the women is poor (Sahle 2016).

Maternal mortality is a death of a mother while pregnant or within forty-two days after the termination of pregnancy, irrespective of the duration of the pregnancy from any difficulties related to pregnancy or its management (Boerma n.d.; Gebremedhin 2018; Hill and Carolina 2001). World health statistics reflected that the burden of maternal mortality due to complications during pregnancy and child-birth is higher in developing countries as compared to developed ones. Globally, each year 500,000 mothers die from complications of pregnancy and childbirth. These high rates of maternal mortality during pregnancy and child birth need a great attention, especially in developing countries (Banjari et al. 2015).

Ethiopia is one of the developing countries in sub-Saharan Africa with a high maternal mortality ratio. In Ethiopia, pregnancy and childbirth related complications are a frightening issue and potentially fatal experience for thousands of women of the reproductive age. Maternal mortality is also high in the Amhara region next to Somalia and Afar regions. According to the Ethiopian Demographic and health survey (EDHS), the maternal mortality ratio was estimated as 676 per 100,000 live births and infant mortality 77 per 1,000 live births (Boerma n.d.). The EDHS report showed that maternal death represents 36% of all deaths among women of reproductive age groups (age 15-49). Which is among the highest in the (Kvåle, Olsen, and Hinderaker 2015).

The healthcare sector is one of the most information-intensive sectors. It can keep medical data at a growing rate daily. The ability to use these data to extract useful information for providing quality healthcare service is a crucial issue. As data volume grows dramatically, data analysis based on manual and statistical methods is becoming inefficient and impractical in many domains, including healthcare sectors.

New techniques and methods are required to discover and analyze new patterns or relationships hidden in large databases (Fayyad et al.1996). Data mining is an automated tools that can intelligently assist in transforming the vast amount of clinical data into useful information (Durairaj and Ranjani 2013). Data mining technique the health professionals, decision-makers, and planners to formulate better strategies and policies that have a significant role in reducing and controlling maternal mortalities (Rani and Govrdhan 2010).

Data mining has tremendous use in the health sector. Healthcare organizations are applying data mining technologies to control costs and improve efficiency. In medical and healthcare areas, massive amounts of data about patients, hospital resources, disease diagnosis are stored (Durairaj and Ranjani 2013; Pushpan and N 2017) are stored. Accordingly, data mining provides a methodology to transform this massive data into useful information for decision making and problem-solving (Mehta and Bhatt 2016). Data mining improves decision making by giving insight into what is happening today and predict what will happen tomorrow.

Thus, data mining is the best approach to extract relevant knowledge for any organization that performs health related activities.

## **1.2. Statement of The Problem**

Improving the health of women is an indication of a strong community. The latest statistics of WHO and UNICEF show that globally, every day nearly 830 women die from pregnancy and child birth related complications. About 99% of these deaths occur in developing countries (Kvåle et al. 2015).

Improving child and maternal health status is one of the Sustainable Development Goals (SDGs) (Press 2017). The SDGs in 2000 set the target to reduce maternal mortality by 75% for world health organization member countries. However, Ethiopia is below the target of SDGs related to maternal mortality rate. Even there was a minimal but insignificant change of maternal death over the last 20 years. Maternal mortality is still high in Ethiopia (Kvåle et al. 2015).

In Ethiopia, the culture of pregnancy follow-up is still very limited. As a result, maternal death due to pregnancy complications is very high. These complications can be direct

obstetric causes or complications and the pre-existing indirect causes such as malaria, HIV/AIDS, anemia, and malnutrition (Mehta and Bhatt 2016).

The underlying research problem that initiates this research issue, to apply data mining techniques using the collected data at hospitals, is the existence of high maternal mortality in Amhara region due to complications during pregnancy and child-birth. Moreover, lack of knowledge model, at the regional-level, that supports health care professionals, planners, and policymakers, to predict maternal mortality and its determinant factors, for maternal mortality reduction is the pushing factor to conduct this research.

Different researches are conducted on the health sector related to maternal mortality. However, previous studies (Berhan and Berhan 2014), (Mekonnen et al. 2016) focused on the identification of determinant factors using statistical methods. This shows there is still a gap using application of data mining techniques for identifying determinant factors of maternal mortality. To fill this gap, this study aims to construct a predictive model for maternal mortality by extracting interesting patterns and knowledge using the data available at the hospitals, particularly the maternal obstetric dataset. To this end, the study attempts to answer the following research questions.

1. What are the major determinant risk factors to be mined using data mining techniques that contribute to the patterns of maternal mortality?
2. Which data mining algorithms are more appropriate to construct maternal mortality predictive models?

### **1.3. Objectives of the Study**

#### **1.3.1. General Objective**

The general objective of the study is to develop a predictive model using data mining techniques that can predict patterns and determinants of maternal mortality in Ethiopia.

### **1.3.2. Specific Objective**

To achieve the general objective, the research has the following specific objectives.

- To identify the major determinant risk factors using data mining techniques that cause for maternal mortality.
- To develop a predictive model for maternal mortality prediction using the clinical dataset.

### **1.4. Scope and Limitation of the Study**

The scope of the study is limited to using data mining techniques to develop a predictive model that is capable of predicting maternal mortality and identify the major determinants of maternal death in Ethiopia using the data taken from Debre Markos referral hospital, Felege Hiwot Specialized referral hospital and Addis Alem hospital which are located in Amhara region.

In this study, classification techniques are used to construct the predictive model that enables to predict maternal mortality patterns. The key attributes used for the selected problem domain are the age of the mother, obstetric complications, mode of delivery, maternal condition, HIV test result and maternal status. The major limitation of this research is, for the developed model an operational prototype is not developed due to time constraints.

### **1.5. Significance of the Study**

The main significance of the study is analyzing the existed dataset and development of the prediction model. This helps physicians, health policymakers, stakeholders, health institution managers (private and public) to be aware of the problems associated with maternal death during pregnancy and they can take corrective measures in controlling the problems associated with maternal health. The study also contributes to identify the determinant parameters from the dataset for effective follow-up and treatment activities in the future and will serve as a base for conducting further investigation in the area of the problem domain. The research result, can be a significant input for health care sectors to provide quality and essential health care services for mothers in the community and health facilities. In general, the society, government, physicians,

policymakers, researcher, domain experts, and hospitals will get benefit from the research result.

## **1.6. Methodology of the Study**

To achieve the objectives of the study and answer the research question the following research methods are used.

### **1.6.1. Research Design**

In this study, design science research framework which is the problem-solving model has its roots in business was adopted for understanding, executing and evaluating this research and this study, follows a six-step hybrid knowledge discovery process (KDP) model (Cios and Kurgan n.d.). The hybrid process model has been chosen since it combines the best features of both the KDD and CRISP-DM (Cross Industry Standard Process) process model. This process model offers more detailed feedback mechanisms, a more general and research-oriented descriptions of the steps (Han and Kamber 2006).

The hybrid process model contains six steps. The initial step is understanding the problem domain. This step defines the problem and learning about the current solution to the problem. Discussion with domain experts and reviewing different kinds of literature that focus on data mining applications and techniques in health care were used as supporting sources. The second step is data understanding, understand the general property of the data (data incompleteness, redundancy, missing value, and noise in data are observed), selection of sample data from the dataset and discussion with domain experts to have general understanding of the data are some activities of this step. Finally, this phase verifies the usefulness of the data with respect to the data mining goals. The next step is data preparation, in data preparation phase data cleaning (such as filling missing values, detecting outliers), data formatting, data transformation was done. In the data mining step, appropriate data mining algorithms were selected and run on the prepared data. Data mining algorithms and techniques were experimented to create predictive models using J48, naïve Bayes, and Support vector machine (SVM).

In evaluation of the discovered knowledge step, using performance evaluation methods, the developed model was evaluated to interpret the knowledge patterns properly. These methods are sensitivity, specificity, accuracy, recall, precision, and 10-folds cross-

validation and percentage split for test option was used to measure the performance of the model. 30% of the dataset were selected on random set selection from the initial raw data to evaluate the accuracy of the developed predictive model. Checking whether the discovered knowledge novel, interesting, and interpretation of the results by domain experts (Rani and Govrdhan 2010) .

### **1.6.2. Literature Review**

In order to have a deep understanding on the problem of this study, it is vital to review different literatures that have been conducted in the field so far. For this reason, related literature such as books, journals, and articles that focus on data mining techniques in health care are consulted to understand the domain knowledge, concepts, and methods that are important for developing the predictive model.

### **1.6.3. Data Collection Technique**

In this study, the required data are acquired from documented sources. These documented sources of data are acquired from the maternal delivery register books and medical journal articles using document analysis technique.

### **1.6.4. Implementation Tool**

RStudio programming language is used to develop the predictive model. The researcher selects RStudio because of the following features: programming and statistical language, has in-built functions for data analysis, support more than 8000 packages, simple and easy to learn, R is also rich in statistical functions which are indispensable for data mining (Sahle 2016).

## **1.7. Organization of the Thesis**

This thesis is organized into six chapters. The first chapter deals with the general overview of the study, including the background of the study, statement of the problem, objectives, scope, and contribution of the research. The second chapter presents the reviewed literatures, which briefly discusses data mining applications in healthcare industries and data mining techniques to be used in the study domain.

The third chapter mainly focuses on the overall research methodology; how the research conducted, tools, algorithms used to develop the models. The fourth chapter

deals with data understanding and data pre-processing. Chapter four also discusses about predictive model developments and performance measures of the developed models, and presents the analysis of the result. The last chapter focused on making conclusions and recommendations to show further research directions.

## **CHAPTER TWO**

### **2. Literature Review**

#### **2.1. Overview of Data Mining**

This chapter presents a review of related literatures on data mining. This includes what data mining is, data mining process models, and a variety of data mining tasks in the healthcare sector.

##### **2.1.1. Data Mining**

Enormous amounts of data across a variety of fields are collected and stored to generate new information. The fast-growing, massive databases, beyond the human ability for comprehension without powerful tools. This abundance of data needs to couple with powerful data analysis tools for making scientific discoveries and uncover valuable hidden patterns (Parali and Bednar n.d.).

Data mining is one of a computational tool that assists humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. Commonly, in the health sector, important decisions are made based on the physician's intuition, not on the information-rich data stored in databases. The main reason behind this is the decision-maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. This is further due to the lack of researches in data mining to mine the hidden patterns through the collected data (Priyadharsini and Thanamani 2014).

Data mining is a computerized method for extracting previously unknown, valid, and actionable information from a large volume of database and enables the organization to predict future trends and behaviors (Fayyad, Piatetsky-shapiro, and Smyth 1996; Han and Kamber 2006). Data mining is a way of searching, analyzing, and scrutinizing a huge amount of data, to discover predictive and descriptive patterns, relationships, and any significant statistical correlations (Luan n.d.; Pushpan and N 2017).

(Priyadharsini and Thanamani 2014) data mining helps organizations to make proactive learning-driven decisions and solve the problems that were traditionally consuming too much time and other resources. Different techniques and algorithms are used to accomplish the tasks of data mining. The figure below shows the connection of data



mining with statistics, Machine learning, Artificial Intelligence, database, and others (Thorat and Kute 2014).

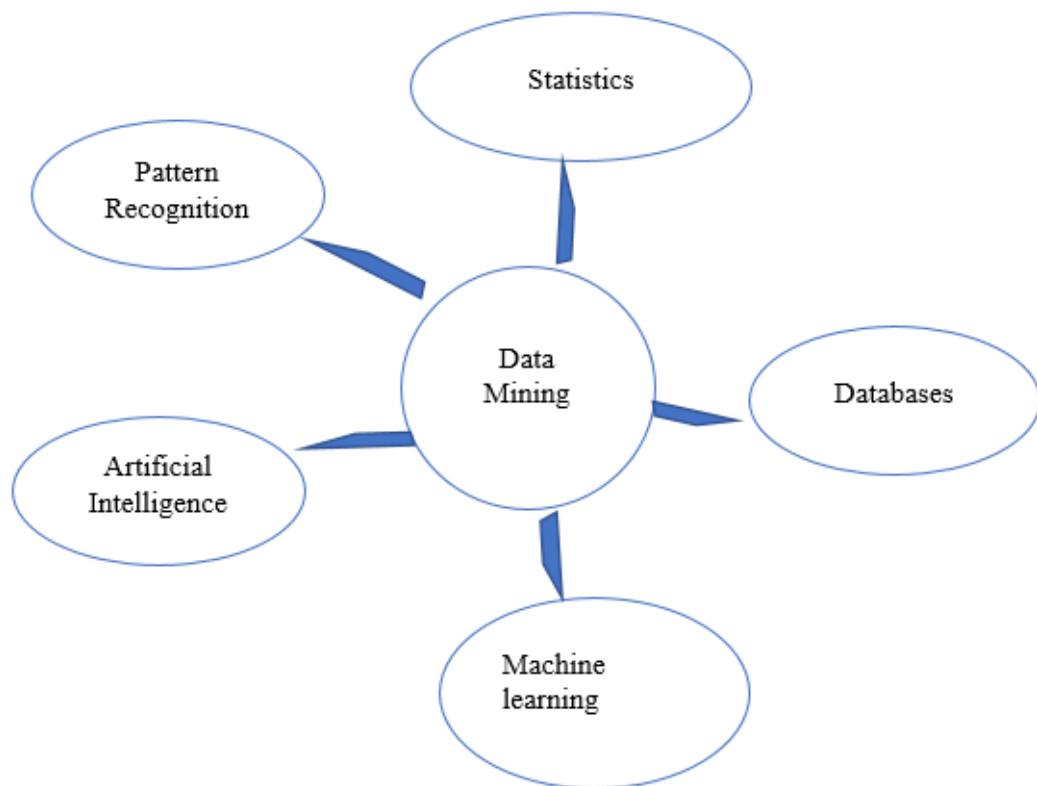


Figure 2.1: Data Mining Architecture with Other interdisciplinary Fields (Han and Kamber 2006)

## 2.2. Data Mining Process Models

There are different types of data mining process model standards that are used in many research domains. Mostly there are four types of data mining process models which, are widely used by the researchers to discover patterns and information from huge datasets. These are Knowledge discovery in databases (KDD), Cross-Industry Standard Process for data mining (CRISP-DM), SEMMA (sample, explore, modify, model, and access) and Hybrid process modes. In this research, the researcher applies a hybrid process model. Which, combines the features of both KDD and CRISP-DM (Luan n.d.; Pushpan and N 2017) (Cios and Kurgan n.d.).

### **2.2.1. Knowledge Discovery in Database (KDD)**

Different scholars deal about data mining and the different phases or steps of the KDD Process. KDD is a multidisciplinary activity to understand, analyses, and extract useful information from a large quantity of data (Fayyad et al.1996). It aimed to give attention that knowledge is the end product of data-driven discovery and it is popular in the artificial intelligence and machine-learning fields.

According to (Priyadharsini and Thanamani 2014) Knowledge Discovery in Database is the automatic extraction of implicit, previously unknown, and useful knowledge from a large volume of data. The KDD processes are interactive and iterative which incorporates many steps with several decisions made by the user (Za, Pole, and Science 1999). the main tasks of knowledge discovery are to extract specific information from previously existing databases and convert it into understandable patterns.

KDD contains a list of iterative and sequential steps as shown in Figure 2.2. The first step is understanding the application domain, which is the initial step in the data mining process. The data mining task starts by clearly understanding the problem domain with relevant prior knowledge. The next step is selecting and creating a target dataset, or focusing on a subset of variables or data samples, on which discovery is to be performed. Next, data cleaning and pre-processing is done on the data such as handling missing values and removal of outliers to obtain consistent data. Then the transformation of the data using dimensionality reduction (feature selection) or attribute transformation (discretization) methods are applied to the pre-processed sample datasets. This step is critical for the success of the entire KDD process. In the next step, the pre-processed and transformed data is changed to meaningful knowledge by applying appropriate data mining techniques such as classification, clustering and regression, and data mining algorithms such as Naïve Bayes, Decision tree, and SVM. Finally, the discovered knowledge based on the mined patterns such as rules, classifications, and predictions is evaluated. Then apply the discovered knowledge to incorporate with another system for further action. The following figure shows the overall processes of KDD.

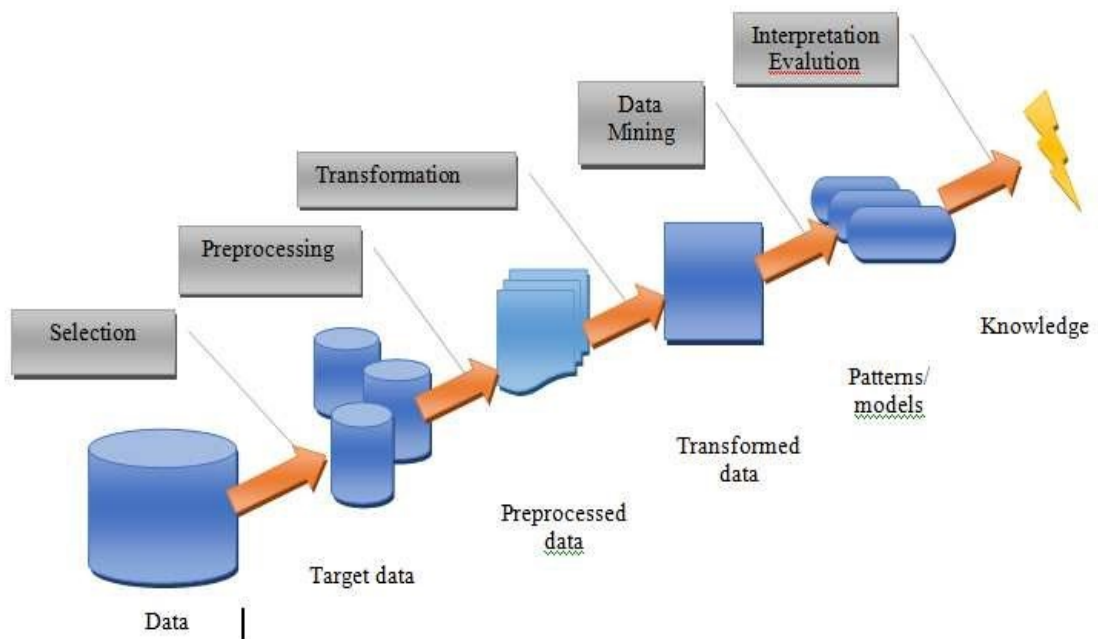


Figure 2.2: The KDD process (Fayyad et al.1996)

### 2.2.2. The CRISP-DM Process

CRISP-DM is a process model for developing general DM and KD projects (Wirth n.d.). The CRISP-DM organizes the data mining process into six phases (i.e. the Business understanding, data understanding, data preparation, modeling, evaluation, and deployment). Each phase helps the organizations to understand the data mining processes.

CRISP-DM is a vendor-independent which means it can be used with any data mining tool and applied to solve any DM problems. To deal with common needs and issues, a group of organizations proposed a reference guide, CRISP-DM (Cross Industry Standard Process for data mining), to develop data mining projects.

The CRISP-DM process model for data mining provides an overview of the life cycle of a data mining project. As depicted in Figure 2.3, the CRISP-DM process model begins with Business Understanding. This phase focuses on the understanding of the domain area from the business point of view. After assessing the current situation, the next task involves converting the business problem into a data mining problem. and then developing a preliminary plan designed to achieve the research objectives. The next phase is Data Understanding which, focuses on collecting initial datasets,

familiarity with the data, describe and explore the data. In this phase, greater emphasis is given for quality and initial understanding of the data. The third step is Data Preparation. The primary goal of this phase is constructing the final dataset from the initial raw data to be used by modeling tools. This stage includes operations such as dimensionality reduction (like feature selection and sampling), data cleaning (like handling missing values), removal of noise, data transformation (like Discretization of numerical attributes). In the model-building phase, selection of modeling techniques and algorithms (such as classification, regression, clustering, and summarization), construction of models, and generation of test designs are going to be applied. During the evaluation phase, the performance and effectiveness of the models created in the previous phase are going to be assessed based on particular metrics. Besides this, performance measurement and analysis are made on the models constructed concerning the goal of the business (Pushpan and N 2017).

Deployment phase, Although The purpose of the model is to increase knowledge gained from the data, and the knowledge gained need to be organized and presented in a way that the user can understand and use it. The following figure shows the CRISP-DM process.

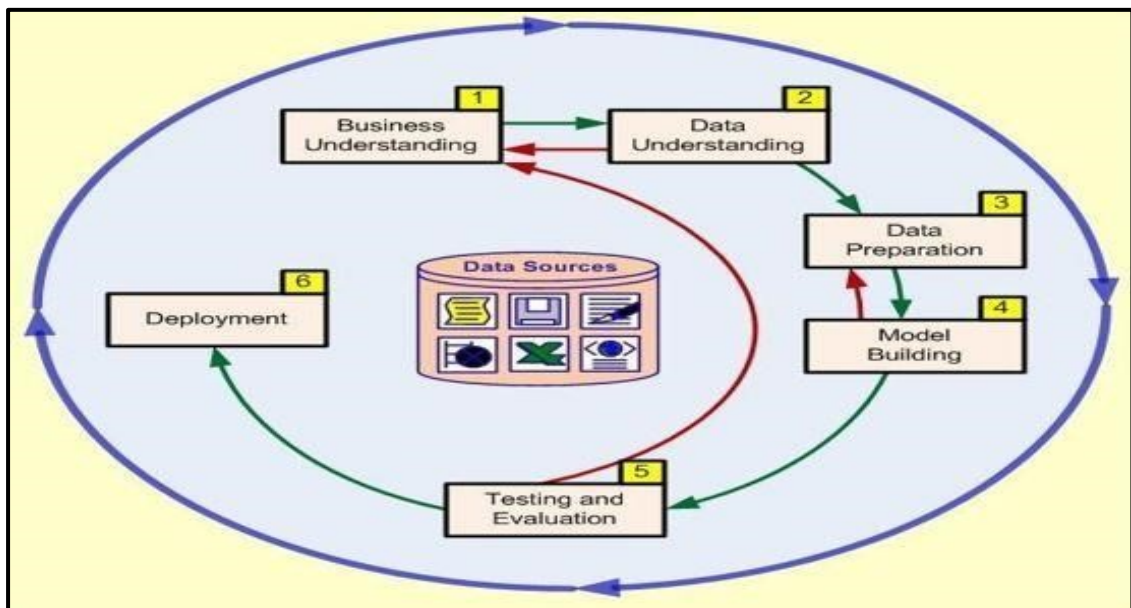


Figure 2.3: The general architecture of CRISP-DM (Wirth, n.d.)

### 2.2.3. SEMMA Process Model

SEMMA is a methodology and approach produced by the SAS institute. The acronym SEMMA-stands for Sample, Explore, Modify, Model, and Assess refers to the core process of conducting a data mining project. Beginning with a statistically representative sample of data, users can apply exploratory statistical and visualization techniques, select and transform the most momentous predictive variables, model the variables to predict outcomes, and check the model's accuracy (Wirth, n.d.) (Za et al. 1999). The SEMMA model divides the data mining into five stages for the process.

- **Sample:** The first step involves sampling the data by extracting a portion of a large quantity of dataset big enough to contain the important information, yet small enough to manipulate quickly.
- **Explore:** This phase involves the exploration of the data by searching speculatively for unforeseen trends and anomalies in order to gain understanding and ideas.
- **Modify:** This phase comprises modification of the data by creating, selecting, and transforming the variables to give emphasis on the model selection process.
- **Model:** This stage involves modeling the data by allowing the software to search automatically for a variable combination that reliably predicts a desired outcome.
- **Assess:** This stage involves the assessment of the data by evaluating the usefulness and reliability of the results gained from the data mining process.

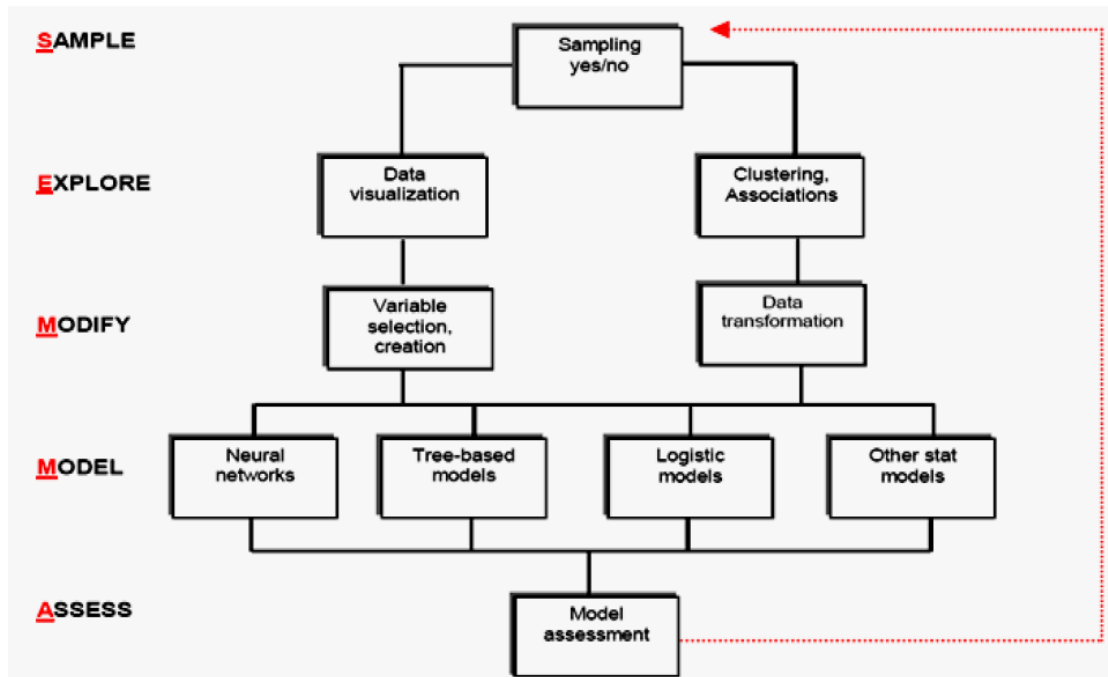


Figure 2.4: SEMMA Process Model (Wirth, n.d.)

#### 2.2.4. Hybrid Model

The hybrid model was followed as a framework to guide the overall activities performed in this study. The development of both the academic model and the industrial model has led to the growth of the hybrid model. It is the combination of aspects of both KDD and CRISP-DM process models. This process model has six basic steps to achieve the overall goals of the data mining process (Cios and Kurgan n.d.). These are Understanding of the problem domain, understanding of the data, preparing the data, data mining, evaluating the discovered knowledge, and finally deploy or use the discovered knowledge for real application in the domain area. The following are descriptions of the six steps of the hybrid process model.

- **Understanding of the problem domain:** In this step one can communicate and closely works with the domain experts to define the problem, review books, documents, magazines, journals, conference papers and others that mainly focuses on data mining techniques and applications in the healthcare domain.it also involves identifying the key participants of the study, determine the research goal, setting solutions to the problem.
- **Understanding of the data:** This step involves collecting the sample data, understanding the data source, and its description. Listing of attributes, checking for completeness, redundancy, missing values, dimensionality reduction, discretization of

numerically continuous attributes to nominal, evaluating the essences of attributes to the research objective is also a major activity undertaken in this step. Finally, data verification is performed based on the usefulness of the data concerning the data mining goals.

- **Preparation of the data:** This is the key step in the data mining process on which the success of the entire knowledge discovery process depends. It usually consumes half of the entire research effort. In this stage, all necessary activities important for data mining are completed. It involves identification of data mining techniques and algorithms, pre-processing the sample data for mining activities, and selecting appropriate data mining tools. Data pre-processing includes data cleaning like checking completeness of the records, removing or correcting for noise or outliers, handling missing values. The cleaned data were further processed by feature selection (reduce its dimensionality), and derive new attributes by discretization. Due to the nature of the datasets classification data mining technique was selected to classify the sample dataset using the algorithms that meets the specific input requirements for the selected DM tools.
- **Data mining:** It is an important research process to know more about the contents of the data and its analysis purpose. Data mining techniques and algorithms are applied in this stage to discover potentially useful, interesting patterns and develop the models. This step involves usage of the planned data mining techniques, tools and selection of the new ones if needed. The data mining tools include many types of algorithms, pre-processing techniques. In the conducted research classification technique is used to develop the model that can solve the specified problems. The training and testing procedure is designed and the model is constructed using the chosen data mining tools and the generated data model is verified in the testing step. The researcher used decision tree (J48), Bayes (naïve Bayes), and SVM classification algorithms and RStudio model development tools.
- **Evaluation of the discovered knowledge:** This step includes understanding the result, checking whether the discovered knowledge is novel, interesting, and the impact of the discovered knowledge, interpretation of the result by domain experts. Only the approved models were retained. The performance of each model developed in the study is measured using accuracy, sensitivity, specificity, recall, and precision. 10-fold cross-validation and percentage split performance evaluation techniques are used to check the performance of the classifier in classifying the dataset.

- **Using the discovered knowledge:** This is the final step; it consists of planning where and how to use the discovered knowledge and determines the success of the entire knowledge discovery process and dissemination of the application area in the current domain to the other domain. The result of the findings in the study would be used by the apprehensive healthcare interested parties and experts. Therefore, interested domain experts and researchers can get access to the research results to support the decision-making process, or use it for further research in the area or any other applicable reasons.

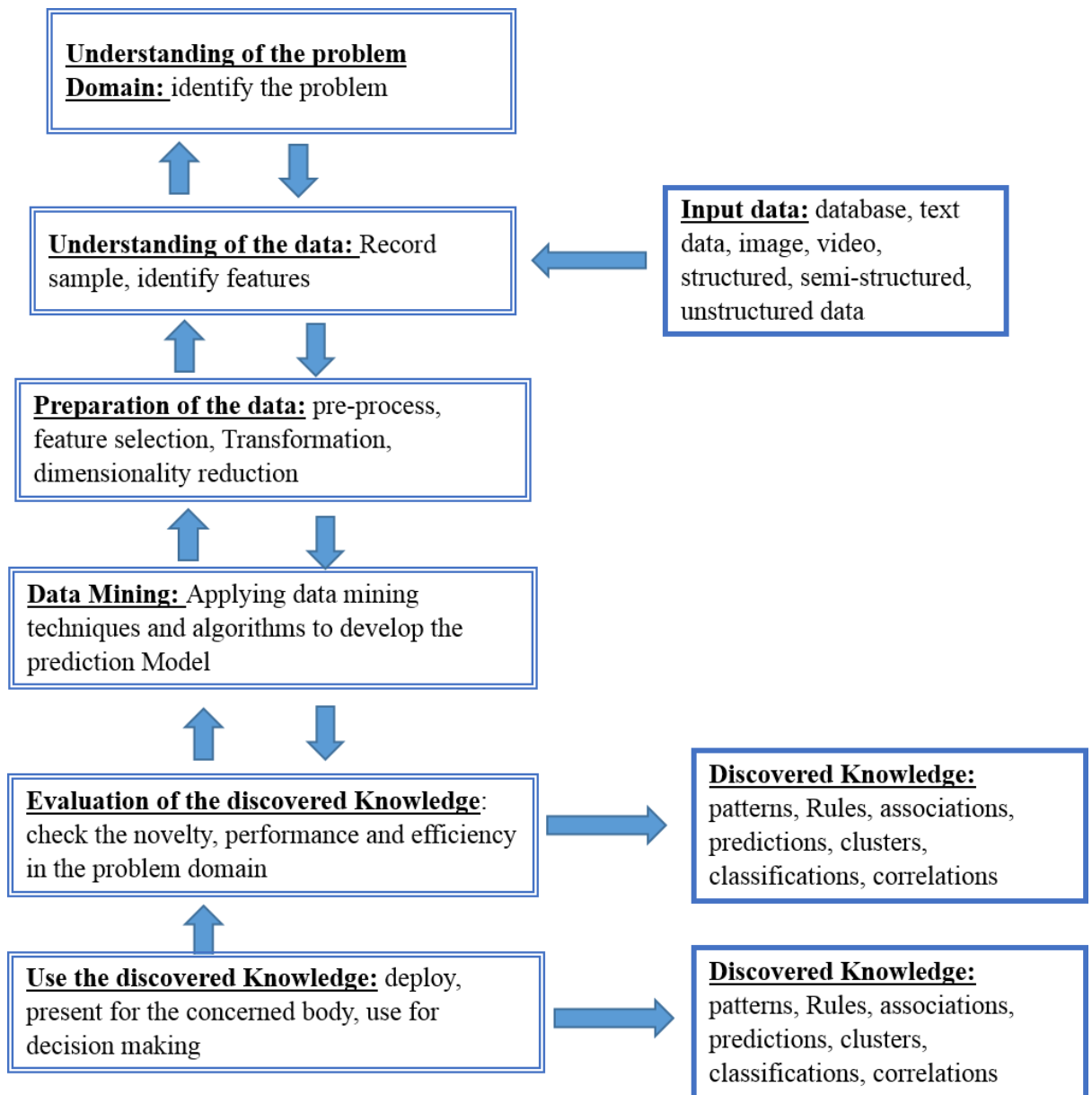


Figure 2.5: The Six steps of Hybrid KDP Model (Cios and Kurgan n.d.)



## **2.3.Data Mining Models and Tasks**

The data mining technique used for modeling purposes is different types depending on the types and nature of the data to be used for the development of the model. A model takes a set of inputs and produces an output. Based on the types of data to be used for analysis and knowledge discovery purposes, data mining models in general, are classified into two main categories (Durairaj and Ranjani 2013). These are predictive and descriptive data mining models. Descriptive (unsupervised) data mining models deal with the general properties of the data in the database, in which there are no known results to guide the algorithms. It simply identifies the pattern and relationship from the data. As shown in Figure 2.6, clustering, association and summarization are some tasks in the descriptive data mining model.

### **2.3.1. Classification**

The two most common predictive modeling tasks are classification and regression. If the label is discrete values, the task is classification and if the label is a continuous value, the task is regression. The focus of this thesis is realizing on predictive models from the knowledge of classification. Classification techniques are the most important approaches for the development of predictive models on the pre-classified cases. In this research, experiments were carried out using three classification algorithms to predict the patterns of maternal mortality in Amhara region based on clinical datasets. These are Decision Tree, Naïve Bayes, and SVM.

Classification is one task of data mining, which contains a set of pre-classified instances to develop a model that can classify a target variable into one of several predefined instances (Priyadharsini and Thanamani 2014). The major objectives of classification techniques are developing an accurate predictive model on the pre-classified target datasets (Za et al. 1999).

The classification process includes learning the data, develop the model, and classifying the new data (Pushpan and N 2017). The learning step takes the input as a training data and it builds a classifier that generates the classification rules. In the classification step, classifying the new data according to the model developed by the training datasets and accuracy of the classifier is tested using test data. If the accuracy

is acceptable, the rules can be applied to new data records. In this study, binary classification was used, which classifies maternal status into alive and died cases.

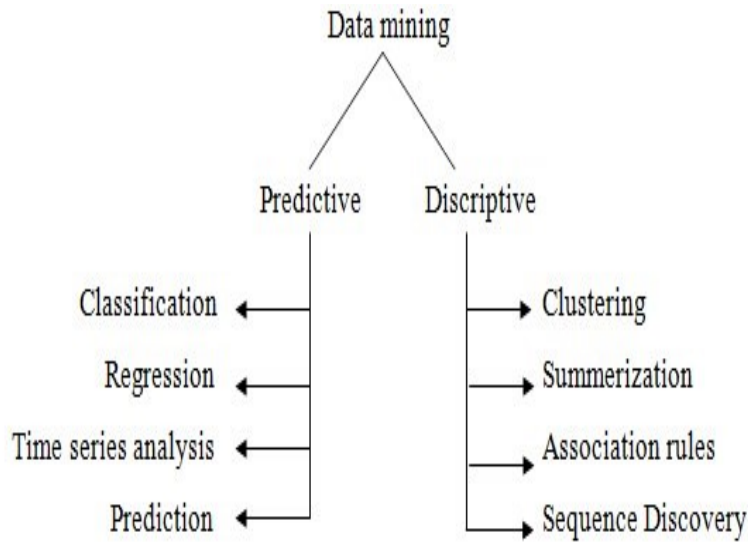


Figure 2.6: Data mining Models and tasks (Fayyad et al., 1996)

#### 2.4. Application of Data Mining Techniques

The field of data mining has been growing rapidly due to its attainments, scientific progress, and broad applicability in various domains (Like education, fraud detection, retail, telecommunication) (Ramageri n.d.).

Education: (Cheng 2017) Suggested that as a multidisciplinary field, data mining is very popular in the education sectors. It gives special attention to analyze educational related data to develop models for improving learners learning experience, enhancing institutional effectiveness, and examining students learning performance.

Retail: Data Mining has its great role in retail industry to identify buying patterns from customers, find associations among customer demographic characteristics, and predict response to mailing campaigns and market basket analysis that lead to improved quality of customer service and good customer retention and satisfaction (Silwattananusarn and Kulthidatuamsuk 2012).

Telecommunication: Data mining in the telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of the resource, multidimensional association and sequential patterns analysis, mobile

telecommunication services, use of visualization tools in telecommunication data analysis and improve quality of service (Keleş 2017).

#### **2.4.1. Application of Data Mining in Healthcare**

The healthcare sector is a complex area in which voluminous and heterogeneous information is generated and collected in a daily basis (Durairaj and Ranjani 2013). An enormous part of this information is found in paper-work. However, making available this information in electronic form and converting it into useful knowledge is not easy work. It needs an expert analysis of their medical data which is time-consuming and tedious for human analysts. The ability to use data in a database to extract useful and meaningful information for quality healthcare service is a key success of healthcare institutions (El-hasnony, Bakry, and Saleh n.d.).

Healthcare institutions have data that contains large quantities of information about the patients, medical conditions, and the parties involved in the institution (Pradhan 2014). The size and complexity of such data are getting higher and higher. Due to this, the healthcare sectors need large storage space and intelligent technologies to retrieve meaningful information from the complex and dirty dataset collections (El-hasnony et al. n.d.). Using traditional methods for extracting meaningful information from the complex datasets is impossible. However, improvements in the fields of statistics, mathematics, machine learning, and data mining allows the extraction of meaningful patterns from large datasets.

Healthcare data mining offers numerous opportunities to explore hidden patterns from healthcare huge datasets (Keleş 2017). These patterns, hidden from these huge datasets of the health sector, provide new medical knowledge. These new medical bits of knowledge are useful for physicians to diagnosis diseases, prognoses treatments and outbreak predictions, estimate the resource use and patient numbers in hospitals, predicting length of stay of patients in hospitals, designing plans for effective information management systems. It is also applicable for classifying the patient data according to factors such as age, gender, race, treatment, to determine the high-risk factors in surgeries (Milovic & Milovic, 2012) (Pradhan, 2014).

Generally, data mining in health sectors serves as a decision support system to ease the life of the physicians. Figure 2.7 showed the overall cycle of data mining in healthcare.

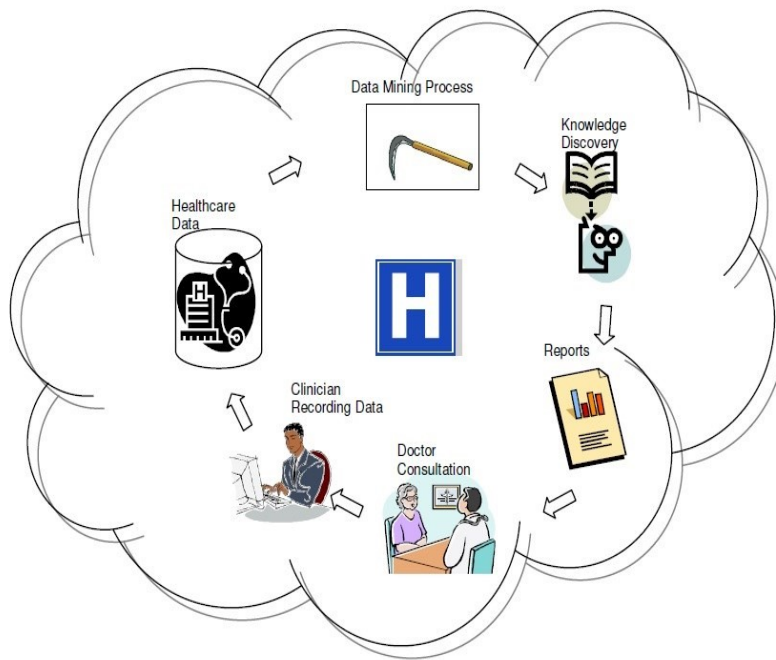


Figure 2.7: Process of data mining in healthcare (Thorat & Kute, 2014)

## 2.5. Related works

To have a deep understanding of the problem domain, it is vital to review several relevant literatures such as books, journal articles, proceeding papers, and other sources from the internet that are related to the study area and data mining tools and techniques. Some of the related works that have been conducted on maternal mortality and its risk factors, the application of data mining in the health care sector are presented below.

(Mekonnen et al. 2016) conducted research on the causes of maternal mortality in Ethiopia between the period 1990 and 2016 and the objective of this review was to document the causes of maternal deaths and risk factors contributing to deaths aggravated by pregnancy and its management in Ethiopia over the period 1990 to 2016. The methodology used was a systematic review with meta-analysis on the causes of maternal death that were published in scientific journals and grey literature, including the compendium of abstracts presented in the series of annual conferences of the Ethiopian Public Health Association. He reviewed a total of 146 articles (134 from online sources and 12 hard copies) that were identified based on their titles and abstracts. According to the research the main direct causes of maternal death in Ethiopia include obstetric complications such as hemorrhage (29.9%; 95% CI: 20.28%-39.56%), obstructed labor/ruptured uterus (22.34%; 95% CI: 15.26%-29.42%), pregnancy-

induced hypertension (16.9%; 95% CI:11.2%-22.6%), puerperal sepsis (14.68%; 95% CI: 10.56%-18.8%), and unsafe abortion (8.6%; 95% CI: 5.0%-12.18%). In recent years, hemorrhage has been the leading cause of mortality, followed by hypertensive disorders of pregnancy and sepsis, while the contributions of obstructed labor and abortion have decreased over the period. The most reported indirect causes of maternal death were anemia (10.39%; 95% CI: 4.79%-15.98%) and malaria (3.55%; 95% CI: 1.50%-3.30%). Finally, He concluded that the nationwide registration of causes of maternal death should be strengthened to understand the causes in detail.

(Berhan and Berhan 2014) conducted research on the causes of maternal mortality in Ethiopia. The methodology used was a systematic review of eighteen health facility-based maternal mortality studies conducted between 1980 and 2012 in Ethiopia. Emphasis was given to the proportion of maternal mortality due to direct causes and their case fatality rates. The findings of the review have shown that the top four causes of maternal mortality in the year 1980-1999 were abortion related complications (31%), obstructed labor/uterine rupture (29%), sepsis/infection (21%), and hemorrhage (12%). however, in the last decade, the top four causes of maternal mortality were obstructed labor/uterine rupture (36%), hemorrhage (22%), hypertensive disorders of pregnancy (19%) and sepsis/infection (13%). Finally, the reviewer concluded that abortion and infection related maternal deaths have declined significantly in the last decade. Obstructed labor continues to be the major cause of maternal deaths; maternal deaths due to hypertensive disorders and hemorrhage showed an increasing trend.

(Of et al. 2015) conducted a study on the assessment of maternal death and factors affecting maternal death surveillance and response system from the period 8 June 2013 to 7 June 2014 and from 8 June 2014 to 9 March 2015. A cross-sectional facility-based study was conducted in nine health facilities of Dire Dawa. The finding of this study has shown that a total of 45 maternal deaths, 247 maternal complications, and 8,857 deliveries were recorded during the two study periods. Maternal mortality ratios for the two periods were 511 and 505 per 100,000 live births in the baseline and implementation period respectively. The direct obstetric causes were responsible for 41 (91%) of the deaths, of which hemorrhage 27%, hypertension during pregnancy 22% and obstructed labour 18% are the leading causes.

(Hailemariam, Meshesha, and Worku 2015) investigated the application of data mining tools and techniques to develop a model that supports the prediction of adult mortality in Ethiopia, particularly Butajira rural health program. WEKA Version 3.6.8 data mining tool was used. The hybrid model that was developed for academic research was followed. Dataset is preprocessed for missing values, outliers and data transformation. Decision tree and Naïve Bayes algorithms were employed to build the predictive model by using a sample dataset of 62,869 records of both alive and died adults through three experiments and six scenarios. In this study as compared to Bayes, the performance of the J48 pruned decision tree reveals 97.2% of accurate results are possible for developing classification rules that can be used for prediction. If no education in family and the person is living in rural highland and lowland, the probability of experiencing adult death is 98.4% and 97.4% respectively with concomitant attributes in the rule generated. The likely chance of adults surviving in completed primary school, completed secondary school, and further education is (98.9%, 99%, 100%) respectively.

(Sahle 2016) conducted the study on Ethiopic maternal care by predicting the key factors that affect postnatal care visit in Ethiopia through data mining techniques. In this study, the researcher used the WEKA tool to build a decision tree (using the J48 algorithm) and rule induction (using JRip algorithm) techniques. The result proves that J48 rule (93.97 % accuracy) is slightly higher than JRip rule (93.93 % accuracy) and places of delivery, the assistance of health delivery professional, prenatal care health professional and age are the determinant factors that affect postnatal care visit.

(Idowu 2018) Developed a Predictive Model for Maternal mortality in Nigeria using Data Mining Technique. The researcher used the Decision tree, MLP and Naïve Bayes algorithm with 10-fold cross validation test mode. The data set contains 200 records of pregnant women for both training and testing the model. WEKA toolkit is used to build the models and the performance of the classification algorithms on the dataset was measured using recall, precision, accuracy, F-measure, True Positive (TP) and False Positive (FP) rates and area under the Receiver Operating Characteristics (ROC) curves. The model was developed using age, birth spacing, parity, gravidity, attendant at ante-natal clinic, financial status, education status or literacy level of the mother attributes. The output attribute is the chances of safe delivery of the mother i.e. High

safe delivery, average safe delivery, and low safe delivery. MLP outperforms the other two.

Generally, the research work reviewed above was tried to address different health care problems on different data sources. Most studies were focused on assessment and evaluation of factors that cause maternal mortality in some selected areas by using statistical methods, without considering various bias effects of the data. For the utilization of relevant information which is hidden in the data, it is obvious that one needs to be engaged in data mining technique since it is efficient to find unrecognized new knowledge and can mine the knowledge rules automatically from the content of data.

Previously there are also, researches that have been carried out using data mining techniques. But, to the knowledge of the researcher, no previous researches have been done to predict maternal mortality by applying data mining techniques in Ethiopia. Therefore, the main aim of this study is to identify the determinant factors of maternal mortality in Ethiopia using data mining techniques.

Thus, this research has a great contribution to generate patterns that help in planning a better strategy and effective decision making for more maternal health promotion plans and programs.

## **CHAPTER THREE**

### **3. Methodology of the Study**

#### **3.1. Research Design**

For this study, the researcher has adopted a hybrid methodology to build a predictive model using data mining techniques.

The study is designed to develop a maternal mortality predictive model based on the clinical dataset, collected from the three hospitals. As discussed in Section 2.2.4 the hybrid model is built through the combination of features of both KDD and CRISP-DM process models.

The researcher has selected this model for the conducted study on the reason that it provides more general, research-oriented descriptions of the steps and detailed feedback loops that are vital to achieve the research objectives. It includes understanding of the problem domain, understanding of the data, preparation of the data, mining the data, evaluating the discovered knowledge, and finally use the discovered knowledge for real applications in the domain area.



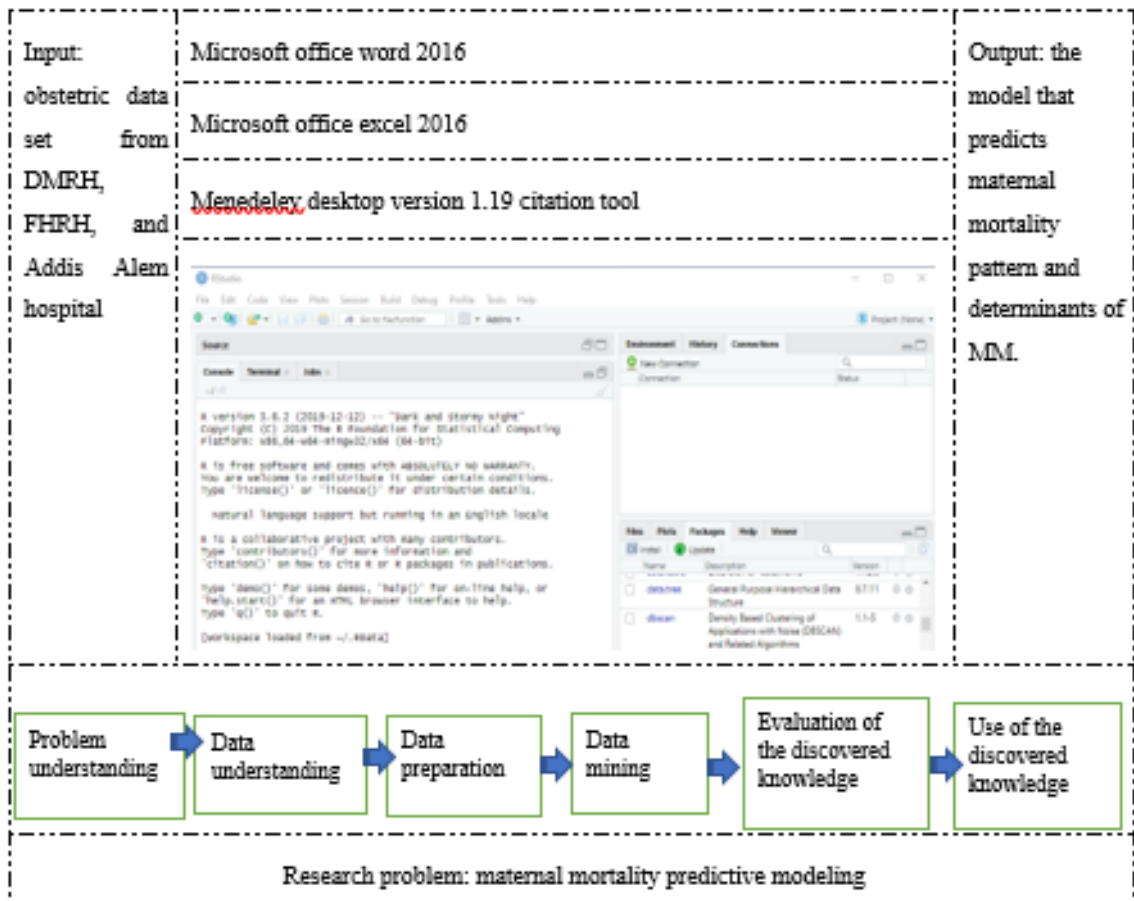


Figure 3.1: Graphical Overview of Overall Research Design and Methodology

In the conducted research predictive (Supervised) data mining model was adopted. The goal of this model is to make predictions of a particular attribute, discover patterns in the data, and to understand the relationships between attributes represented by the data. Predictive modeling lets the value of one variable to be predicted from the known value of the other variable. As depicted in Figure 2.6, Classification, regression, estimation, and prediction are some techniques in predictive modelling to extract useful information from the data. Predictive modelling is a process used in predictive analytics to create a model for future behavior. Predictive analytics is the area of data mining that is focused on forecasting trends. A predictive model is made up of predictor variables that are likely to influence future probabilities. The basic steps to build a predictive model were described as follows.

1. The model is trained using the pre classified data in a subset of a model set. In this step, the data mining algorithm finds the pattern of predictive value.

2. The model is refined using another subset called the test set to prevent the model from memorizing the training set, thereby ensuring that the model is more general and will work better on unseen data.
3. Estimate the performance of the model and compare the performance of several models.
4. The model is applied to the score set.

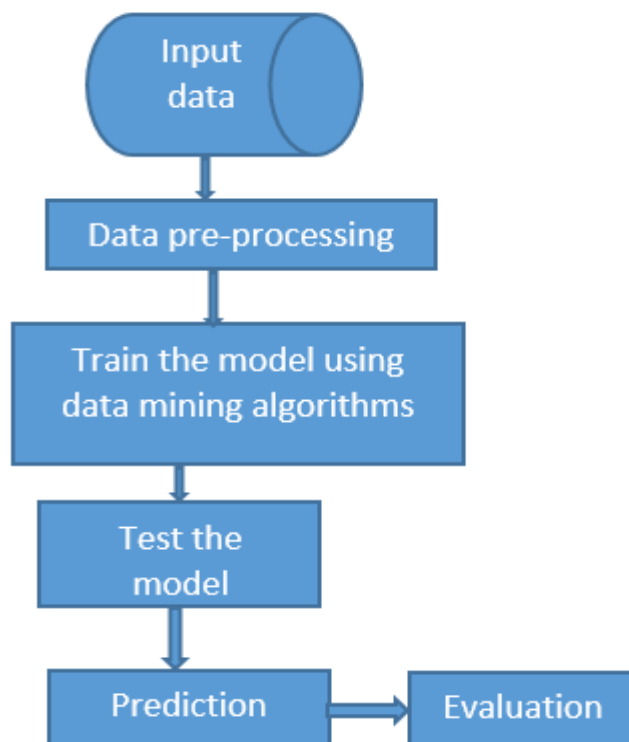


Figure 3.2: Architecture of predictive model for maternal mortality

Once the predictive model is developed using the sample dataset, the model should be checked how it will perform for the data that has not been seen during the model building process.

### 3.2. Tools

An important task performed before building the model was selecting the relevant software that supports the required data mining algorithms. Data mining uses many free open source tools such as rapid miner, geo-miner, WEKA, I-miner, ML knowledge studio, orange, and R studio. For the conducted study the researcher is used the RStudio machine learning tool.

The tool used to analyse the dataset using the selected data mining algorithms. It contains important packages and functions for loading the dataset, classification, regression, clustering, association rule, prediction, and visualization. The researcher selects RStudio because of the following features: programming and statistical language, has in-built functions for data analysis, support more than 8000 packages, simple and easy to learn, R is also rich in statistical functions which are indispensable for data mining. Besides the above features, R is an important tool because of its:

**Performance:** The ability to handle a variety of data sources in an efficient manner.

**Functionality:** The inclusion of a variety of capabilities, techniques, and methodologies for data mining. This software functionality helps to assess how well the tool will adapt to different data mining problem domains.

**Usability:** The usability, applicability by different levels and types of users without loss of functionality or usefulness. In addition to RStudio the researcher used, documentation tools such as MS-word, MS-excel, and Mendeley desktop for citation of references.

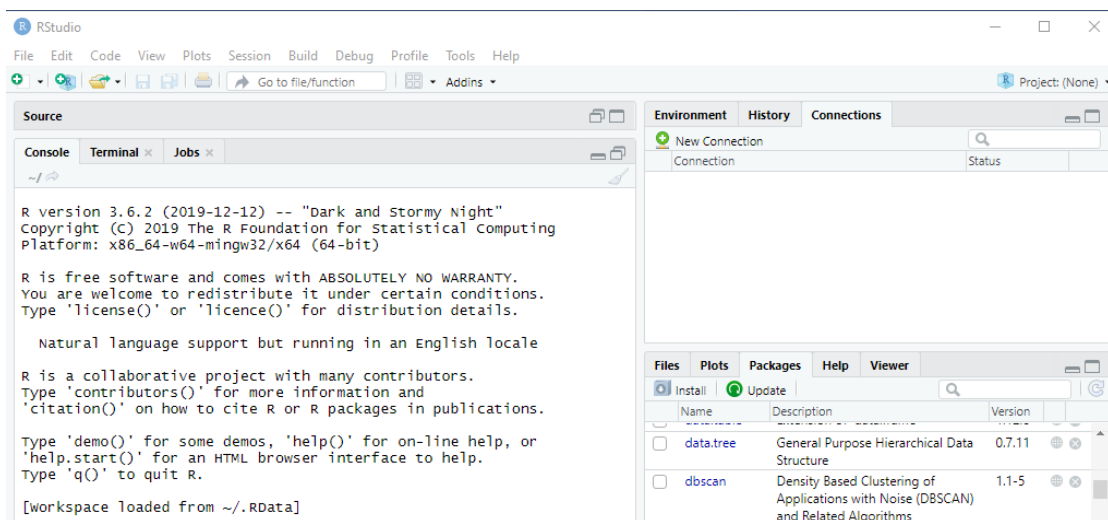


Figure 3.3: RStudio user interface

### 3.3. Algorithms for Model Building

The purpose of building models is to use the predictions for making more knowledgeable decisions. In this study, the researcher has adopted three classification algorithms namely, decision tree by applying J48 classifier, Naïve Bayes, and SVM classifier to develop the planned prediction model and evaluate the accuracy of the model

based on accuracy, sensitivity, specificity, recall and precision. Each algorithm used in this study is described in the following subsections.

### **3.3.1. Decision Tree Classifier**

The decision tree is a powerful method for modeling classification and prediction in data mining (Danilo 2010). It represents a procedure for classifying categorical and numerical data using the selected attributes and rules which can be easily understood by humans and used in the knowledge system.

A decision tree classifier uses a divide and conquers method to split the problem search space into subsets. It is expressed as a recursive partition of the instance space (Rokach and Maimon n.d.). In data mining, a decision tree is a predictive model that can be used to represent classification and regression. Classification trees used to predict a categorical variable, because they place instances in classes or categories for example, alive or died.

There are several algorithms that are based on decision trees. Some of the most common types of algorithms based on decision trees are C4.5, PART and CART (Kim and Loh 2014). For the conducted study Rweka:: J48 decision tree classifier and Gain ratio as splitting criteria. The splitting stops when the number of instances is below a certain threshold value.

#### **3.3.1.1. Decision tree building**

Decision tree is a flowchart-like tree structure model (Han and Kamber 2006), where each internal node (non-leaf node) indicates a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) clutches a class label (Sharma and Kumar 2016). The tree is built by dividing the large dataset into several smaller sets (Danilo 2010).

Decision trees are constructed through recursive partitioning, an iterative process of splitting the training data into one or more sub-partitions until a stopping criterion is met. Each sub partition dominantly consists of examples of one class (Berry and Linoff 2004).

The topmost node in a tree is the root node. Each internal node in a decision tree splits the instance into two or more sub-spaces based on discrete functions of the input attribute values. While the decision tree is constructed, it is possible to generate the rule

to apply it for new cases. The following algorithms show how a decision tree algorithm generates a tree from the given training data. (Han and Kamber 2006)

**Algorithm:** Generate a decision tree from the training tuples of data partition,  $D$ .

**Input:**

- Data partition,  $D$ , which is a set of training tuples and their associated class labels;
- attribute list, the set of candidate attributes;
- Attribute selection method, a procedure to determine the splitting criterion that “best” partitions the data tuples into individual classes.

**Output:** A decision tree.

**Method:**

1. Create a node  $N$ ;
2. If tuples in  $D$ , are all of the same class,  $C$  then
3. Return  $N$  as a leaf node labeled with the class  $C$ ;
4. If the attribute list is empty then
5. Return  $N$  as a leaf node labeled with the majority class in  $D$ ;
6. Apply Attribute selection method ( $D$ , attribute list) to find the best splitting criterion;
7. Label node  $N$  with splitting criterion;
8. If the splitting attribute is discrete-valued and multiway splits allowed then
9. Attribute list  $\leftarrow$  attribute list – splitting attribute; // remove the splitting attribute
10. For each outcome  $j$  of splitting criterion // partition the tuples and grow subtrees for each partition
11. Let  $D_j$ , be the set of data tuples in  $D$  satisfying outcome  $j$ ; // a partition
12. If  $D_j$ , is empty then
13. Attach a leaf labeled with the majority class in  $D$  to node  $N$ ;
14. Else attach the node returned by Generate decision tree ( $D_j$ , attribute list) to node  $N$ ; end for
15. Return  $N$ ;

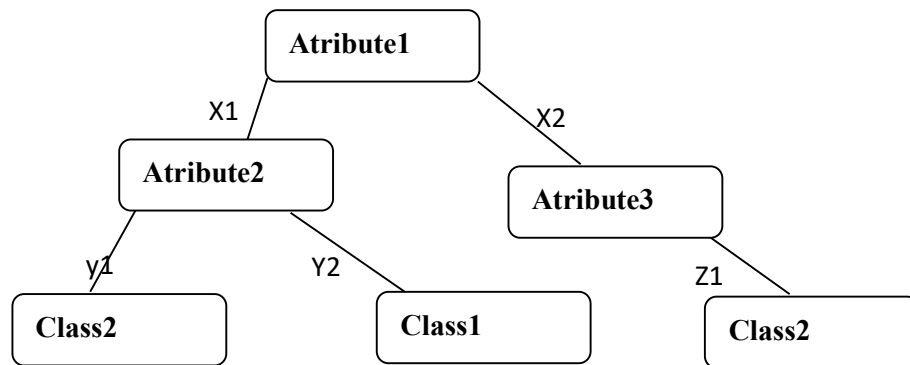


Figure 3.4: A simple decision tree for two class classification (Han and Kamber 2006)

From the above simple decision tree, the following rules can be generated.

Rule 1: If Attribute1= X1 and Attribute2= Y1 then classification =Class2

Rule 2: If Attribute1= X1 and Attribute2= Y2 then classification =Class1

Rule 3: If Attribute1= X2 and Attribute3= Z1 then classification =Class2

### 3.3.1.2. Pruning decision tree

Growing the tree beyond a certain level of complexity leads to overfitting. Tree pruning is needed to avoid a large tree or to prevent overfitting when classifying new data. Pruning refers to reducing the size of the tree that is too large and deeper (Rokach and Maimon n.d.), which increases the efficiency and accuracy of classification.

According to (Patel 2012) pruning methods are categorized into pre-pruning and post-pruning. In post pruning (sometimes called backward pruning) method first built the complete tree and then reduction of non-significant branches and levels of the tree is done. Post-pruning avoids branches from a fully grown tree. In this case, the nodes are pruned by removing its branches. Pre-pruning (forward pruning) prevents the generation of non-significant branches and involves checking whether the tree is overfitting and tries to decide during the tree building process when to stop developing sub-trees. In this study, the Pre-pruning method was used.

### 3.3.1.3. Rule induction

As the name indicates that, rule induction generates rules of some type. IF-THEN prediction rules are a very popular rule in data mining, they are able to represent a discovered knowledge at a high level of abstraction. The learned decision tree is

represented by IF-THEN rules to improve human readability of the decision tree. Each rule can be created from root to leaf node (the leaf node represents the class prediction) and each branch from a leaf node corresponds to the possible values of the attributes.

The rule antecedent (the IF part) contains one or more conditions about the value of predictor attributes whereas the rule consequent (THEN part) contains a prediction about the value of a target attribute. An accurate prediction of the target attribute will improve decision-making process. In the idea of classification, rules are of the form: IF set of conditions then class. In this study, the C4.5 pruned decision tree model is used to induce rules.

### **3.3.2. Naïve Bayes Classifier**

Naïve Bayes classifier is a probabilistic and statistical data mining method with conditional independence assumptions. That means the presence or absence of a particular feature of a class is unrelated to the presence or absence of other features (Leung n.d.).

Naïve Bayes classifier also called simple Bayes and independence Bayes is the easiest model to construct. Because it does not need any complicated iterative parameter estimation scheme. The method is designed for use in supervised learning tasks, in which the task of the learner is to predict the correct class for the test instances, where the training set contains instances for the class variable or attribute. The Bayesian classifier works based on Bayes theorem.

It is a method of classification that does not use rules, unlike a decision tree. Rather, it uses the branch of mathematics known as probability theory to find the most likely of the possible classifications. The naïve Bayes algorithm gives us a way of combining the prior probability and conditional probabilities in a single formula, which the researcher used to calculate the probability of each of the possible classifications in turn. Having done this the researcher chooses the classification with the largest value. Taking into account the nature of the underlying probability model, the Naïve Bayes classifier can be trained very efficiently in a supervised learning setting, working much better in many complex real-world situations.

The a-prior probabilities are prior probabilities for each class in naïve Bayes' theorem. That is how frequently each level of class occurs in the training dataset. The conditional

probabilities are calculated for each variable. it is the likelihoods in naïve Bayes theorem.

Naïve Bayes classifier, assigns a posterior probability to a class based on its prior probability and likelihood given the training data. It computes the maximum posterior hypothesis or maximum likelihood hypothesis (Keller n.d.).

Naïve Bayesian classifier is the most straightforward and widely tested method used for probabilistic induction. This model represents each class with a single probability instant (Langleyflamingostanfordedu et al. 1993) (Hickey 2013).

### 3.3.2.1. Naïve Bayes Theorem

More specifically, as mentioned above Bayes theorem is one main concept which needs to be considered under Naïve Bayesian classifier. Naïve Bayes theorem gives the conditional probability of an event H given on another event X has occurred.

Let  $X = \{X_1, X_2, X_3, \dots, X_n\}$  be a sample, in Bayesian terms which is considered “evidence” and component denotes values made on a set of n attributes. Let H be some hypothesis, such that the data sample X belongs to a specific class C . For classification problems, the main issue is to determine  $P(H|X)$ , the probability that the hypothesis H holds given the “evidence”, (i.e. the observed data tuple X) (Keller n.d.).

$P(H|X)$  is the posterior probability, of H conditioned on X that is, the posterior probability of class given predictor (attribute).  $P(H)$  is the prior probability of the class, or a prior probability of H. Similarly,  $P(X|H)$  is the likelihood which is the probability of the predictor given class and  $P(X)$  is the prior probability of X.

Thus, according to Bayes theorem the required posterior probability,  $P(H|X)$  can be computed from  $P(H)$ ,  $P(X|H)$  and  $P(X)$  probabilities, using the following expressions.

$$P\left(\frac{H}{X}\right) = P\frac{\left(\frac{X}{H}\right)P(H)}{P(X)} \quad eqn(1)$$

### 3.3.3. Support Vector Machine Classifier

Support vector machines (SVMs) are supervised learning methods that can be used for building both classification and regression models. SVMs are an effective method for binary classification to categorize the input data and regression for the estimation of the desired output. For classification, nonlinear kernel functions are often used to transform the input data into a high dimensional feature space in which the input data



becomes more separable compared to the original input space. The support vector machine algorithm works based on the concept of decision planes, called hyperplanes used to classify the values of the target variable.

There can be several hyperplanes in support vector machines that can divide the data points without any error. However, the best hyperplane is the one which has the maximum distance from the nearest points of the classes. The support vector machine tries to find a hyperplane that maximizes the margin distance while minimizing misclassification errors. Then, maximum-margin hyperplanes are constructed to optimally separate the classes in the labeled training data. The nearest data points from the hyperplane that maximizes the distance are support vectors (Meyer 2019).

### **3.4. Model Evaluation Techniques**

After accomplishing classification model creation, the next task is comparing the predictive accuracy of the classifiers for unknown tuples to evaluate the performance of predictive modeling. Accuracy tells us how frequently instances of particular classes are correctly classified as an actual class or misclassified as some other classes.

Performance evaluation methods are chosen to evaluate and interpret the interesting patterns properly. These methods are sensitivity, specificity, accuracy, recall, and precision. 10 folds cross-validation and 70% percentage split; test option also used to measure the performance of the model. 30% of the dataset were selected on a random set selection from the initial raw data to evaluate the accuracy of the developed predictive model.

#### **3.4.1. Methods of Training and Testing the Model**

In every case, the model is built through the training set and perform prediction on the test set. To minimize the problems associated with the random sampling of training and testing data in comparing the predictive accuracy of two or more models, k-fold cross-validation is the best choice. In k-fold cross-validation, the complete data set is randomly split into k mutually exclusive subsets of approximately equal size. In this research, experiments have been done by splitting the dataset into training and testing sets, using percentage split test mode and 10-fold cross validation. The sample code used to split the dataset was annexed in appendix 1.

### 3.4.2. Performance Evaluation of Classification Models

After creating the intended model, comparing the predictive accuracy of the classifier for unknown instances are very important to evaluate the performance of the predictive model.

#### 3.4.2.1. Confusion Matrix

In the field of machine learning, a Confusion Matrix, also known as a contingency table, is a specific table layout that allows visualization of the performance of an algorithm. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The confusion matrix is the summary of prediction results on the classification problem. It is the body of table with  $m$  by  $m$  (row and column) matrix the row corresponds to correct classification and the column corresponds to the predicted classifications. An entry,  $CM_{i,j}$  in the first  $m$  rows and  $m$  columns indicate the number of tuples of class that was labeled by the classifier as class  $j$ . For a classifier to have good accuracy, ideally, most of the tuples would be represented along the diagonal of the confusion matrix with the rest of the entries being closed to zero. The caret package in RStudio contains the function `confusionMatrix()` to provide a confusion matrix and the associated statistics using the predicted outcomes as well as the actual outcomes for the classifiers. confusion matrix holds classifier evaluation metrics like accuracy, error rate, sensitivity and specificity, precision, recall, and F-measure. Table 3.1 shows two class classification confusion matrixes that contain both predicted and actual classes.

Table 3.1:Confusion matrix of two class classification result

Actual class	Predicted class	
	A	B
A	True positive	False Negative
B	False positive	True Negative

**Key:** TN= True Negative, TP =True Positive, FN =False Negative, FP =False Positive

Here are some of the performance evaluation computational techniques on the confusion matrix that are used in this study.

Accuracy is the percentage of correctly classifies instances out of all instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

**Sensitivity (True Positive Rate):** proportion of actual positives which are predicted positive.

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

**True Negative Rate (specificity)** or Recall for False class proportion of actual negatives which are predicted negative.

$$\text{specificity} = \frac{TN}{TN + FP}$$

**Precision (positive predictive value)** proportion of predicted positives which are actual positives.

$$\text{Precision} = \frac{TP}{TP + FP} \text{ -----For True Class}$$

$$\text{Precision} = \frac{TN}{TN + FN} \text{ -----For False Class}$$

**Recall:** proportion of actual positives which are predicted positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

**F- measure:** harmonic mean of precision and recall.

$$\text{F-Measure} = \frac{2(\text{Precision} * \text{recall})}{(\text{Precision} + \text{Recall})}$$

$$\text{Error rate} = \frac{FP + FN}{TP + TN + FP + FN}$$

Error rate of the classifier determines how much percent error is committed by the model which is usually computed as the difference of one and accuracy.

## **CHAPTER FOUR**

### **4. Data Understanding and Pre processing**

#### **4.1. Data Understanding**

Data mining needs to realize the core critical issues during the study. These are, formulate the problem you are trying to solve and use the right data to explore the hidden patterns.

The data understanding phase starts with an initial data collection process for the study and proceeds to other activities to be familiar with the data such as identify data quality problems, detect the appropriate subset of the data being collected, clearly define and have a good understanding of our data to be used for data mining tasks. During data understanding, the initial data collection process involves selecting the representative section of the data which is likely suitable and reliable to meet the objectives stated by applying data collection techniques.

In this research, the data collection process was initially carried out by converting the paper-based format data from maternal delivery and death registration book which is found in the hospitals into a computer or electronic format (usually in a spreadsheet) to take advantage of easier data manipulation and provide compatible interaction with the selected tool.

Table 4.1: list of Attributes in the initial dataset and its descriptions

NO.	Attribute	Description	Data Type
1.	S.N	Serial no. of the mother	Numeric
2.	MRN	The medical record number of the mother	Numeric
3.	Name	Name of the pregnant or delivered mother	Text
4.	Address	The particular address of the mother	Text
5.	Age	The age of the mother during childbirth	Numeric
6.	Date of delivery		Date
7.	Mode of delivery	The method of delivery of the mother	Nominal
8.	Maternal condition	The delivery status of the mother	Nominal
9.	Obstetric complication	The type of complication that faces a mother during delivery	Nominal
10.	HIV test result	The HIV assessment results of the mother	Nominal
11.	Maternal Status	The final status of a mother during pregnancy and delivery	Nominal
12.	New born birth outcome	The final status of the new born baby	Nominal

Each record in the sample dataset corresponds to a single person/patients history which, initially contains personal information, such as S.N, MRN, first name, middle name, last name of the mother, address, age and date of delivery, mode of delivery, maternal condition, HIV test result, maternal status, Obstetric complication, and New-born birth outcome. From this recorded information, the researcher has selected attributes which, help to develop a predictive model for maternal mortality prediction in hospitals.

The dataset of this research initially contains 5017 instances and 12 attributes. Then pre-processing techniques were applied to make it appropriate for the mining process. through discussion with domain experts and support of literatures, from the total of 12 attributes; in consultation with domain experts 5 irrelevant attributes (S.N, MRN, name of the mother, address, and date of delivery) were removed and based on gain ratio values one least important attribute (the new born birth outcome) removed.

#### **4.1.1. Data Collection Method**

For this study, the data was collected from Debre Markos Referral Hospital, Felege Hiwot Specialized Referral Hospital and Addis Alem hospital covering the period 2007 to 2010 E.C. The information collected from the obstetrics and gynecology Center of the hospitals were registered on the maternal delivery register book which is found in the form of papers. These data contain detailed information about maternal conditions, mode of delivery, age, HIV Test result, maternal status, and Obstetric complications. Therefore, analyses were made based on the data available at the obstetrics and gynecology Center of the hospitals. Interview with the concerned body (such as health professionals or domain experts) who have an idea about the study area and review of related documents and manuals were also applied to gather relevant information for the conducted study.

#### **4.2. Data Preprocessing**

Data may have quality problems that need to be addressed before applying any data mining techniques. Data in the real world are highly susceptible to noise, inconsistency, and incompleteness, this is because the data may have a huge size and is obtained from multiple, heterogeneous sources. Data pre-processing is an important step in data mining process which, makes data more suitable for data mining and used to improve the quality of the data (Du n.d.), thus helping to improve the accuracy and efficiency of the overall data mining results (Han and Kamber 2003).

The collected data about the maternal delivery situation, which was employed for this study, suffer from different constraints, for example, missing value and noise data. These constraints result difficulties in performing the predefined data mining objectives and tasks. To develop an optimal model, a cleaned and automated dataset is needed.

Thus, the main purpose of data pre-processing is to transform the dataset so that their information content is best palatable to the mining tool. It also allows the miner to produce faster and better models. The figure below shows the major tasks that were done during the data pre-processing stage.

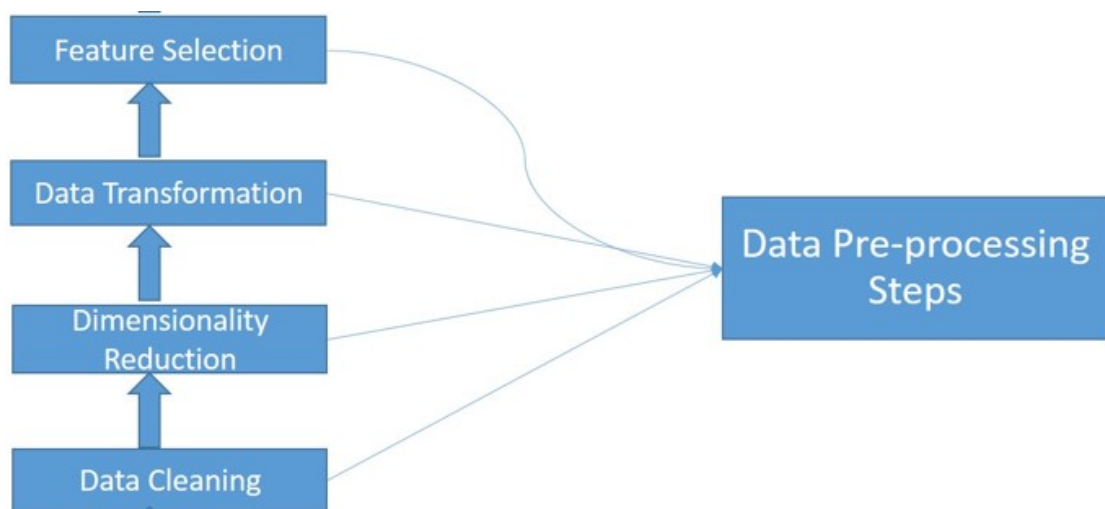


Figure 4.1: Data pre-processing steps for quality data mining

#### **4.2.1. Data Cleaning**

Data cleaning is a process of detecting and removing errors and inconsistencies from the data to improve the quality of the data. Data cleaning is a routine task, which involves filling missing values, smooth, noise, remove outliers, and resolving inconsistencies to improve data quality. The correctness of an output of the data mining results highly depends on the cleanness of the data. Hence, data cleaning reveals the correction of data quality problems.

Missing or insufficient attributes and values are examples of data quality problems that may confuse data analysis tasks such as learning and hinders the performance of algorithms. handling of missing values and noise removal are some activities performed in the data cleaning stage of this study.

##### **Handling of missing value**

Missing data may occur due to equipment malfunction, inconsistent with other recorded data, misunderstanding, or certain data may not be considered as important at the time of entry. Missing values may lead to the difficulties of extracting useful information from the dataset. Therefore, identifying and solving the problem of missing data is given a high priority in the field of data mining and knowledge discovery.

When there are missing values, instead of leaving them as missing, there are several methods that can be used for filling these missing values.

Handling missing values by appropriate methods does not affect the quality of data.

Some of the methods used for filling up missing values include:

- Filling the missing values manually-manually search for all missing values and fill them with the appropriate values. Mostly, this is done when the missing values to be filled are known.
- Ignoring the records when an entered tuple is empty- in this method, the missing value is completely ignored as the analysis is carried on.
- Using a global constant-replace all the missing values by some constant such as a label like “unknown” or “?”
- Using the mean/mode imputation method- replace missing data with their mean (numeric attribute) and mode (categorical attribute) of all cases observed. This is the most common method of filling the values quickly.
- Case deletion method – this method omits those instances with missing data and analyses the remaining instances.

The missing values may be due to recorder problems at the time of encoding the data from paper to Excel format or it might be due to physician’s problem when they fill the history of the mother. To handle the missing (Not available) values in this study, the researcher used mode for factor attributes. NA in RStudio represents the missing value. This technique replaces the missing value with the most frequented values.

Table 4.2: list of attributes with missing and estimated values

Attribute name	Attribute type	Percentage of missing value	New estimated value	Technique applied
Mode of delivery	factor	0.001%	SVD	mode
obstetric complication	factor	0.001%	PPH	mode



## **Noise Removal**

Noise occurs due to random error or variance in a measured variable. There may be incorrect attribute values in the data due to data entry problems, data transformation problems, duplicate records, incomplete and inconsistent data. In the study dataset, mode of delivery attribute consists of 34 SVCSD and 24 SD noise values, respectively, due to encoder's error. Thus, the researcher has discussed with domain experts and data encoders. Finally, the noise or inconsistent attribute values were removed from the dataset.

### **4.2.2. Data Transformation**

Data transformation is a technique of transforming or consolidating the data into forms appropriate for the mining process. It may include the following strategies such as Smoothing, Aggregation, Generalization, Normalization, discretization, and feature construction. To increase the quality of data in this study the researcher used the discretization technique on numeric attributes to minimize distinct values of the attribute.

#### **4.2.2.1. Data Discretization**

Discretization is one of a data pre-processing technique used to reduce the number of values for a continuous variable by grouping them into a number of intervals or a smaller number of distinct ranges (Witten n.d.). And give a label for each interval (Ryan 2006).

In this research, the mother's Age attribute was discretized to reduce the unlike values of the attribute to obtain knowledge and to make the dataset suitable for mining tools. Then this attribute was discretized into eight labels using equal-width intervals binning to make the continuous value attributes valuable for mining purposes and to interpret the model easily. The process in equal width (distance) involves finding values as maximum (max) and minimum (min). Equal width discretization is a simple discretization method that divides the range of observed values for a feature into  $k$  equal-sized bins using the formula  $\text{interval} = (\text{max} - \text{min}) / k$ . The following table shows the discretized labels of the mother's age.

Table 4.3: Data discretization

Age Binned label (Year)	Frequency	Percentage
Less than 15	22	0.28
15-20	784	1.1
20-25	2492	33
25-30	2904	38
30-35	964	13
35-40	411	0.53
40-45	60	0.08
>45	10	0.01

#### 4.2.3. Selecting the Attributes

Most machine learning algorithms are designed to learn the most appropriate attributes for making useful decisions. Attribute selection is an essential component (Setiono n.d.). It is the process of selecting a subset of relevant attributes based on certain criteria for model creation (Guyon 2003). Attributes selection technique is more important to get the minimum best attributes for prediction. Attributes that are highly relevant for developing the predictive model were selected and others that are not relevant to the specified objectives were removed.

Real-world data usually contains imperfect, irrelevant, and redundant features to mining tasks. Therefore, removing these features through the feature selection method may reduce storage and computational costs and improves the learning performances of the algorithm(Li et al. 2016; Oreski and Novosel 2014).

Attributes in the initial dataset were selected based on different criteria's including its relevancy to the research objectives. Therefore, in this research attributes were selected using gain ratio value and with the help of the domain experts. Taking all the attributes kept in the original dataset and feeding them into the data mining tool faces problems. It takes too much time to build a model when the number of variables is increased and produces an incorrect model when there is an extraneous column.

Besides, the ideas gained from domain experts and literatures, the researcher evaluated the information content of the attributes using the select attribute techniques from the

RStudio data mining tool. RStudio provides different attribute selection mechanism through coding and by installing the necessary packages for each technique.

To select the top-ranked attributes from the total list of attributes in the sample dataset the gain ratio attribute evaluator method was used using the [GainRatioAttributeEval\(\)](#) function in RStudio.

Gain ratio attribute evaluator evaluates the value of an attribute by measuring the gain ratio with respect to the class and rank all attributes according to their Gain value.

The attributes with the maximum gain ratio are selected as the splitting attribute. With this regard, attributes selected using gain ratio technique in RStudio are maternal condition, Obstetric complication, Age, HIV Test Result, Maternal status, and Mode of delivery.

The Gain ratio for the given attribute A is defined as

$$\text{GainRatio}(A) = \text{Gain}(A)/\text{SplitInfo}(A) \quad \text{eqn (2)}$$

The split information value represents the potential information generated by splitting the training data set D into V partitions, corresponding to V outcomes on attribute A.

Table 4.4: List of selected attributes and their descriptions

NO.	Attribute	Description	Data Type
1.	Obstetric complication	The type of complication that faces a mother during delivery	Factor
2.	Maternal condition	The delivery status of the mother	Factor
3.	Mode of delivery	The type of delivery of a mother	Factor
4.	Age	Age of the mother during pregnancy	Factor
5.	HIV test result	The HIV test result of the mother at child birth	Factor
6.	Maternal status	The final status of the mother	Factor

In RStudio the variable is nominal by making it a factor. The factor in R stores the nominal values as a vector of integers in the range (1.....k) where k is the number of unique values in the nominal variable and an internal vector of character strings (the original value mapped to these integers). The sample code used to select the importance attribute is annexed in appendix 2.

Table 4.5: List of selected attributes based on their gain ratio value

Attribute name	Attribute importance/rank value
mcondition	0.52414244
complication	0.49938191
Age	0.08076433
htr	0.07918931
mod	0.05011313

#### 4.2.4. Data Formatting

While the researcher selects the implementation tool for developing the required predictive model, the next task is making the available data format compatible with the RStudio tool.

First of all, the researcher encodes the patient's history from paper format into Ms-excel file format (.xls) then into comma-separated value (CSV) to create R understandable file format for experimentation.

```
'data.frame': 7647 obs. of 6 variables:
 $ Age      : Factor w/ 8 levels "'\]'(-inf-15]'\'",...: 5 2 7 3 5 3 4 3 3 5 ...
 $ mod      : Factor w/ 4 levels "CS","FORCEPS",...: 4 4 1 4 4 2 4 1 4 4 ...
 $ mcondition : Factor w/ 2 levels "STABLE","UNSTABLE": 1 1 1 1 1 1 1 1 1 1 ...
 $ htr      : Factor w/ 2 levels "NR","R": 1 1 1 2 1 2 1 1 1 2 ...
 $ complication: Factor w/ 8 levels "APH","ECLAMPSIA",...: 2 4 4 4 4 1 4 4 4 4 ...
 $ mstatus   : Factor w/ 2 levels "ALIVE","DIED": 2 1 1 1 1 2 1 1 1 1 ...
```

Figure 4.2: CSV files used for maternal mortality prediction

### **4.3. Experimentation and Discussion**

#### **4.3.1. Model Building**

To build the model, the first task performed is importing the cleaned and pruned dataset into the RStudio software. The data organized as CSV format as depicted in Figure 4.2 was provided for the RStudio tool.

The test option used in this study were K- fold (K=10) cross-validation and 70/30% percentage split. In 10-fold cross-validation, the total sample dataset is partitioned into 10 parts or 10 folds. Where K is the number of splits to make in the dataset. K-fold cross-validation was employed for randomly sampling the training and test data samples. These validation methods are used to check the performance of the model through k-times. Subsequently, within each iteration, a different fold of the data is held out for testing the model and the remaining K-1 folds are used for learning (or training) the classifier.

The essence of using 10 -fold cross-validation is, its ability to perform extensive tests in many datasets with different learning techniques and 10 is the right number of folds in K-fold cross-validation to get the best estimate of error. 10 -fold cross-validation is also used to reduce the bias associated with the random sampling of the training and holdout data samples by repeating the experiment 10 times, each time using a separate portion of the data as holdout sample.

After pre-processing is done on the data, a total of 4959 instances with 6 attributes are ready to build a model using a Decision tree, Naïve Bayes, and Support vector machine algorithms. Those algorithms were selected since they are easy for interpretation and understanding. The selected attributes during data pre-processing or preparation phases are age, obstetric complication, maternal condition, mode of delivery, HIV test result and, maternal status.

The 4959 instances (cases) represent both alive and died classes. From this instance, 4831 (97%) cases comprise alive class and the rest 128 (2.6 %) are died, class. This shows that there is a class imbalance that leads to ignoring the minority class by the classification model. That is, if there is a class imbalance, the classifier might bias to the majority class. According to (Chawla et al. n.d.), if such a condition happens, different techniques needed to balance the classes. Therefore, to balance the target

attribute values, the researcher used synthetic Minority oversampling technique (SMOTE) which is an operation where the minority class is oversampled by generating the synthetic examples of minority class and adding them to the dataset.

By applying the SMOTE technique, the classification accuracy of the minority class has been improved by a certain level. Therefore, through SMOTE a total of 7647 instances (4831 alive and 2816 died) were found, and the subsequent experimentations were conducted based on this sample dataset.

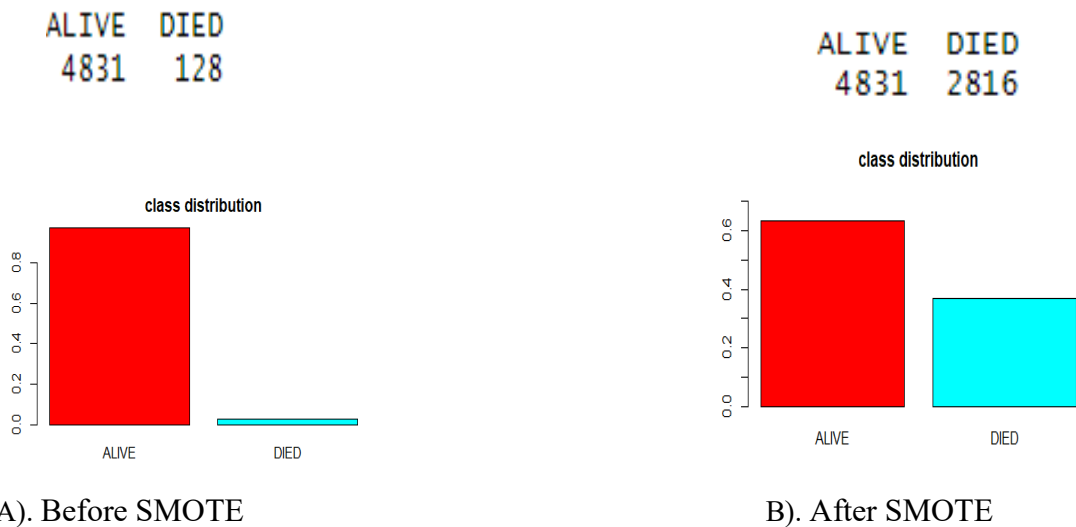


Figure 4.3: Side by side view of the class variable using SMOTE

For classification, by making use of RStudio 1.2.5, a total of ten experiments using different schemes were carried out. eight of the experiments were conducted for constructing a decision tree, one experiment was for the Naïve Bayes classifier, and the remaining one for SVM model generation using percentage split test mode and 10-fold cross-validation.

Table 4.6 : Experiments and schemes

Experiments	Schemes
J48 Unpruned tree model generation	J48 -U -M 2
	J48 -C 0.4 -M 3
	J48 -C 0.5 -M 2
J48 Pruned tree model generation	J48 -C 0.25 -M 5
	Naïve Bayes
Naïve Bayes classifier	Naïve Bayes

Support vector machines	SVM radial base function
-------------------------	--------------------------

After the modeling tool is selected and performance evaluation criteria are set, building a model with several parameters that could manage the model generation process would be the next task.

**4.3.1.1. Model Building Using J48 Algorithm**

R programming and statistical language provide many built-in functions and packages for both regression and classification model development. To develop the J48 model in RStudio the researcher used Rweka package and `RWeka::J48()` function. J48 is one of the classification algorithms in RStudio to develop a decision tree model. J48 decision tree classifier supports both numeric and nominal predictors and nominal class variable values.

An attempt was made to find a better model by changing the important parameters of the J48 classifier. These parameters allow greater control of the user in the process of learning the models. The confidence factor/threshold parameter used to set a limit so that the algorithm makes more or less pruning. The confidence factor is important to produce optimal and accurate decision tree with high correctly classified instances. The MinNumObj option represents the minimum number of instances per leaf. This option allows us to dictate the minimum number of instances that can constitute a leaf.

The most basic parameter is the tree pruning parameter. Depending on how the training and testing data have been defined, the performance of an unpruned tree may superficially appear better than a pruned tree. This can be the result of overfitting. Pruning allows fewer, more easily interpreted results. More importantly, pruning can be used as a tool to correct potential overfitting problems. It is important to develop models by intelligently adjusting these parameters.



Table 4.7: different parameters and its value for J48 decision model building

<b>Experiments</b>	<b>Parameters</b>			
	Pruned	Confidence factor	minNumobj	Test mode
<b>Experiment#1</b>	True	0.4	3	70% percentage split
<b>Experiment#2</b>	True	0.5	2	70% percentage split
<b>Experiment#3</b>	True	0.25	5	70% percentage split
<b>Experiment#4</b>	False		2	70% percentage split
<b>Experiment#5</b>	True	0.4	3	10-fold cross validation
<b>Experiment#6</b>	True	0.5	2	10-fold cross validation
<b>Experiment#7</b>	True	0.25	5	10-fold cross validation
<b>Experiment#8</b>	False		2	10-fold cross validation

Table 4.8: Experimental results of J48 decision tree classifier

Performance measurements	Experiments							
	#1	#2	#3	#4	#5	#6	#7	#8
Accuracy (%)	97.52	<b>97.56</b>	97.43	97.5	97	97.43	96.95	97
Sensitivity (%)	99.29	<b>99.29</b>	96.55	96.76	96.9	97.4	96.9	0.971
Specificity (%)	94.67	<b>94.78</b>	98.93	98.7	97.8	97.8	98	97.9
Precision	96.76	<b>96.83</b>	99.36	99.22	97	97.5	97	0.972
Recall	99.29	<b>99.29</b>	96.55	96.76	96.9	97.4	96.9	0.971

As one can see from Table 4.8 above the researcher conducted different classification experiments by changing important parameters of the decision tree (J48) classifier. To select the best predictive model, the models were evaluated by their accuracy, sensitivity, specificity, precision, and recall.

From the table above we observed that among eight experiments, the first four experiments are carried out using 70/30 percentage split test mode. Here, higher accuracy (97.56%) is obtained by pruned J48 decision tree classifier with confidence factor=0.5, minNumobj=2. And the next, four experiments are implemented using 10-fold cross-validation test mode. Here, also pruned J48 decision tree classifier with confidence factor=0.5, minNumobj=2, provides higher accuracy (**97.43%**). This shows that classifier with the same parameter provides higher classification accuracy in both test modes.

As observed from the table above the eight experiments performed nearly equally well with the highest accuracy score of 97.56% obtained by pruned J48 decision tree classifier with confidence factor=0.5, minNumobj=2 and with percentage split test option and the lowest accuracy score is 96.95% obtained by pruned J48 decision tree classifier with minNumobi=5 and 10-fold cross-validation.

An essential feature of J48 is its ability to generate outputs both in tree forms and rule sets. The classifier with higher classification accuracy is selected to generate tree and rule to achieve the intended objective.

Therefore, from the eight experiments carried out above, the researcher selected a pruned J48 decision tree classifier (with a classification accuracy of (97.56) implemented on experiment #2 with a 70/30% percentage split test mode to build a predictive model.

From the experiments, it can be concluded that although it significantly pruned the size of the tree, as the minNumObj parameter value increased, the accuracy of the classification algorithm decreased. This is true in both test modes or options of classification. The reason for this is record in a given leaf could be in different classes and there could be attributes that could further split the records in the same node into disjoint classes. The sample code used to develop the pruned J48 model is annexed in appendix 3.

Name	Type	Value
resultJ48	list [6] (S3: J48, Weka_tree, Weka_	List of length 6
classifier	S4 (rJava:jobjRef)	S4 object of class jobjRef
predictions	factor	Factor with 2 levels: "ALIVE", "DIED"
call	language	J48(formula = mstatus ~ ., data = train, control = Weka_control(M = 2, C = ...
handlers	list [1]	List of length 1

Figure 4.4: Pruned J48 classifier detailed in Rweka interface

#### 4.3.1.2. Model Building Using Naïve Bayes Algorithm

The researcher tried to show the experiment on the naïve Bayes algorithm to get the best fitted model for the classification and prediction of the maternal status and risk factors for MM based on clinical datasets. This experiment was designed to explore the performance of the Naïve Bayes model using different test modes. The package e1071 contains a function named `naiveBayes()`, which is used to perform Bayes classification.

Table 4.9: Experimental results of Naive Bayes classifier

Model	Performance measurement	Test option	
		70% percentage split	10-fold cross validation
Naïve Bayes	Accuracy (%)	<b>94.2</b>	93.97
	Sensitivity (%)	<b>95.86</b>	94
	Specificity (%)	<b>91.36</b>	93
	Precision (%)	<b>95</b>	94
	Recall	<b>95.86</b>	94

As shown in the above experiment table, the Naïve Bayes model with a 70/30% percentage split test option performs better classification accuracy scores of **94.2%**, sensitivity and precision are high as well. The sample code used to develop the naïve Bayes model is annexed in appendix 6.

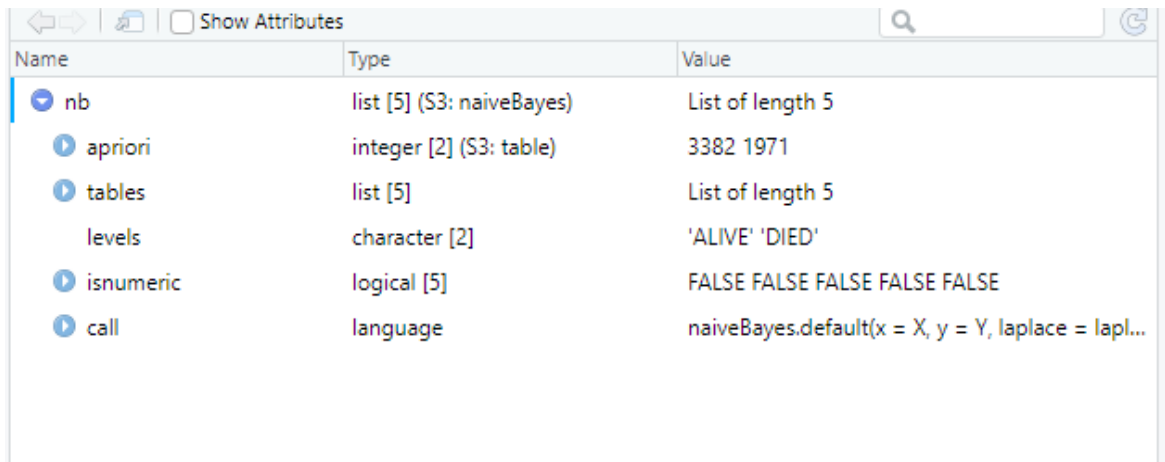


Figure 4.5: Naive Bayes classifier detailed

#### 4.3.1.3. Model Building Using Support Vector Machine Algorithm

The main purpose of this experiment is to evaluate the performance of the SVM classifier using radial basis function as a kernel property. The package e1071 also contains an `SVM()` function which is important to develop the SVMs classification model.

Table 4.10: Experimental results of SVM classifier

model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Recall	Test mode
SVM radial	<b>96.64</b>	<b>94.69</b>	<b>100</b>	<b>100</b>	<b>94.69</b>	<b>70% percentage split</b>
	96	96.1	97.7	96.4	96.1	10-fold cross validation

As one can observe from the table above, the two experiments were conducted using SVM radial function, the highest classification accuracy **96.64%** is obtained by 70% percentage split test option. The sample code used to develop the SVM model is presented in appendix 7.



obtained by the models and how the model meets the business domain to classify the maternal outcome as Alive and Died.

Table 4.11 : Performance summary of the three Models

Model	Accuracy	sensitivity	specificity	Precision	Recall
Decision tree model (J48 pruned)	97.56	99.29	94.78	96.83	99.29
Naïve model Bayes	94.2	95.86	91.36	95	95.86
SVM radial	96.64	94.69	100	100	94.69

As observed from Table 4.11, the J48 pruned tree model with confidence factor 0.5 has higher classification accuracy **97.56%** than the SVM radial classifier which has the second-highest accuracy (96.64%) and the naïve Bayes model has the lowest accuracy (94.2%) using 70/30 percentage split test mode. J48 pruned model with confidence factor 0.5 also achieves good sensitivity, precision and recall. Therefore, the J48 pruned model with confidence factor 0.5 and 70% is the best model to predict maternal mortality. The performance of J48 classifier is better because of parameter difference from SVM radial and Naïve Bayes.

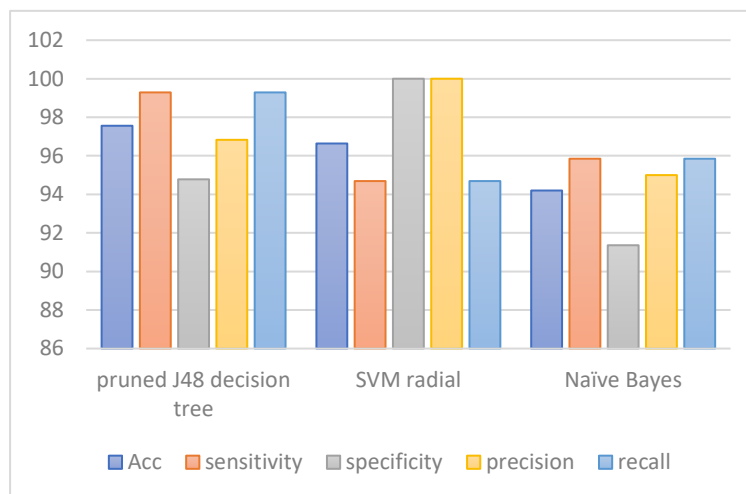


Figure 4.7 : visualization of performance comparison the three models

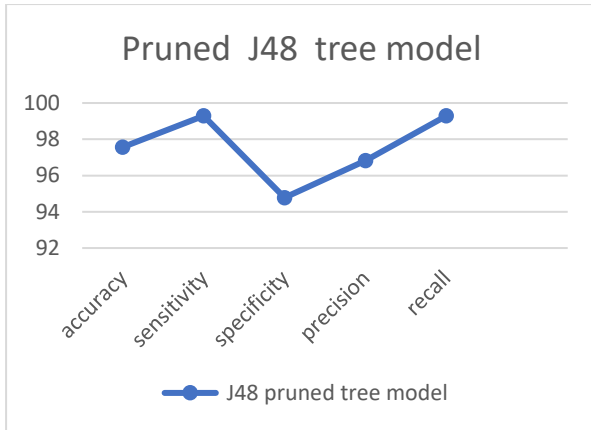


Figure 4.8: Performance measures of the selected model

Table 4.12 : Confusion Matrix result of the selected J48 -C 0.5 -M 2

		Predicted Class		
		Alive	Died	Total
Actual class	Alive	1403	10	1413
	Died	46	835	881
	Total	1449	845	2294

The Confusion Matrix for J48 pruned decision tree shown in Table 4.12 depicts that out of the total 1413 records the model correctly classified 1403(99.29%) records in the class of Alive whereas 10 (11%) of the records were incorrectly classified as Died while actually, they were in the Alive class. The model correctly classified 835 (94.78%) records in the class of died out of 881 records that in fact died and the remaining 46 (3.1%) records were misclassified to the Alive class in fact they were in the Died class. This indicates that from the total records, 2238 (97.56%) records were correctly classified while the remaining 56 (2.44%) records were classified incorrectly. Hence, this showed that records whose class Alive were classified with a minimum error as compared to records of the Died class.

### 4.3.3. Rule Extraction

In the conducted research, rules are generated by using J48 pruned decision tree classifier based on its accuracy. Decision tree is that the best-known method for deriving rules from classification trees. Each path from the root node to the leaf can be transformed into an IF- THEN rule to improve human readability of the decision tree.



If the condition is satisfied, the conclusion follows. Nine strong rules that cover most of the points in the study were selected. The numbers which appear in the parentheses next to the class label indicate that the number of correctly classified and incorrectly classified records, respectively.

The importance of the rules, attributes used to construct the rules and the association of the attributes with the predicted class, predicted by the rules were evaluated based on the comments given by domain experts and reports of previous studies.

**Rule1.** If mcondition = UNSTABLE) and complication = PPH and mod =CS and Age 20-25 **Then** maternal status=DIED (321.0/2.0)

The first rule is selected from the rules generated by J48 pruned decision tree model gives correct result of 321 out of 323 instances that it covers. This rule indicates that the likelihood of the mother to die is about 99.7%. This rule also predicts that PPH is one key factor (complication) which causes a mother to die during pregnancy.

**Rule2.** If mcondition = UNSTABLE) and mod = SVD and Age = 20-25 and complication = PPH **Then** maternal status=DIED (83.0/11.0)

The second rule also selected from the rules generated by J48 pruned decision tree that gives correct result of 83 out of 94 cases that it covers. Its success fraction is 88 %. Based on this rule a pregnant woman at age between 20 and 25 and have PPH complication are highly in danger to die.

**Rule3.** If mcondition =STABLE and complication =PPH and Age = 25-30 **THEN** maternal status=DIED (156.0/18.0)

This rule is selected from the J48 pruned decision tree which gave the correct result of 156 out of 174 instances that it covers. This rule shows that the likelihood of the mother to die at age between 25 and 30 by the complication factor PPH is about 89.7%.

**Rule4.** If mcondition = UNSTABLE and complication = ECLAMPSIA and mod= CS **Then** maternal status=DIED (115.0/1.0)

This is another rule selected from the J48 pruned decision tree which gives the correct result of 115 out of 116 instances that it covers. This rule shows that the likelihood of predictability of the mother to die by the complication factor eclampsia is about 99%.

**Rule5. If** mcondition = UNSTABLE and Age = 25-30 and complication = PPH **THEN** maternal status=DIED (673.0/16.0)

This is another rule selected from the J48 pruned decision tree which gave the correct result of 673 out of 689 instances that it covers. This rule shows that the likelihood of the mother to die at age between 25 and 30 and by the complication factor PPH is about 97.7%.

**Rule6. If** mcondition= UNSTABLE and complication = APH and mod=OTHER and Age = 30-35 **THEN** maternal status=DIED (166.0)

The rule states that a mother having complication factor APH at the age between 30 and 35 has 100% chance likely to die.

**Rule7. If** mcondition = STABLE and complication = ECLAMPSIA and Age = 30-35 **THEN** maternal status=DIED (56.0)

This is another rule selected from the J48 pruned decision tree which gave the correct result of 56 out of 56 instances that it covers. This rule shows that the likelihood of the mother to die at age between 30 and 35 and by the complication factor eclampsia is about 100%.

**Rule8. If** mcondition = STABLE and complication = APH and mod = SVD and Age=20-25 **THEN** maternal status=DIED (62.0/2.0)

The rule is selected from the rules generated by J48 pruned decision tree model gives correct result of 62 out of 64 instances that it covers. This rule indicates that the likelihood of the mother to die is about 97%. This rule also predicts that APH is another key factor (complication) which causes a mother to die during pregnancy.

**Rule9. If** maternal condition = STABLE

Complication = NO: **THEN** ALIVE (3124.0)

This rule gives a correct result of 3124 out of 3124 instances that it covers and the rule indicates that a mother at stable condition and with no complication has 100% probability to alive. Domain experts confirmed that mother with no pregnancy complication has high chance to survive.

Thus, as the discovered knowledge shows, the attributes obstetric complication, maternal condition and woman age were found to be the major determinant factors for maternal mortality.

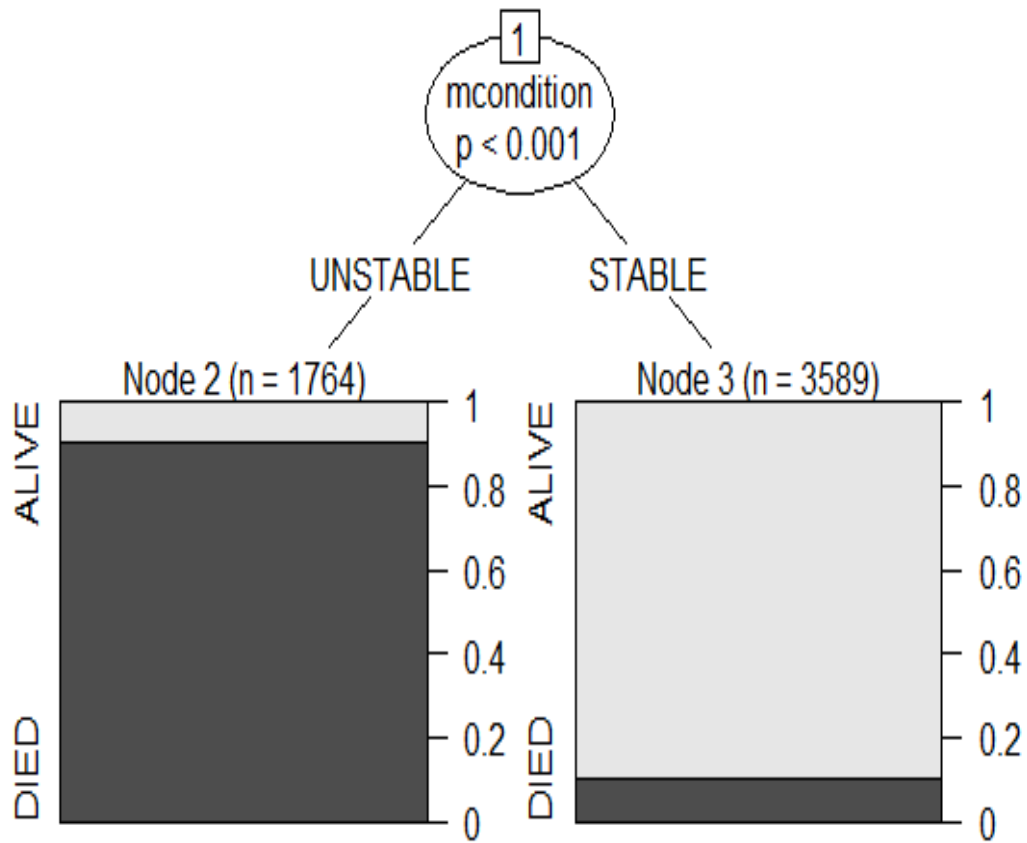


Figure 4.9: Predictive relationship between maternal status and maternal condition

As one can see from Figure 4.9 above, maternal mortality is high when the mother is unstable during her pregnancy or delivery time. Thus, maternal status is highly correlated with the maternal condition.

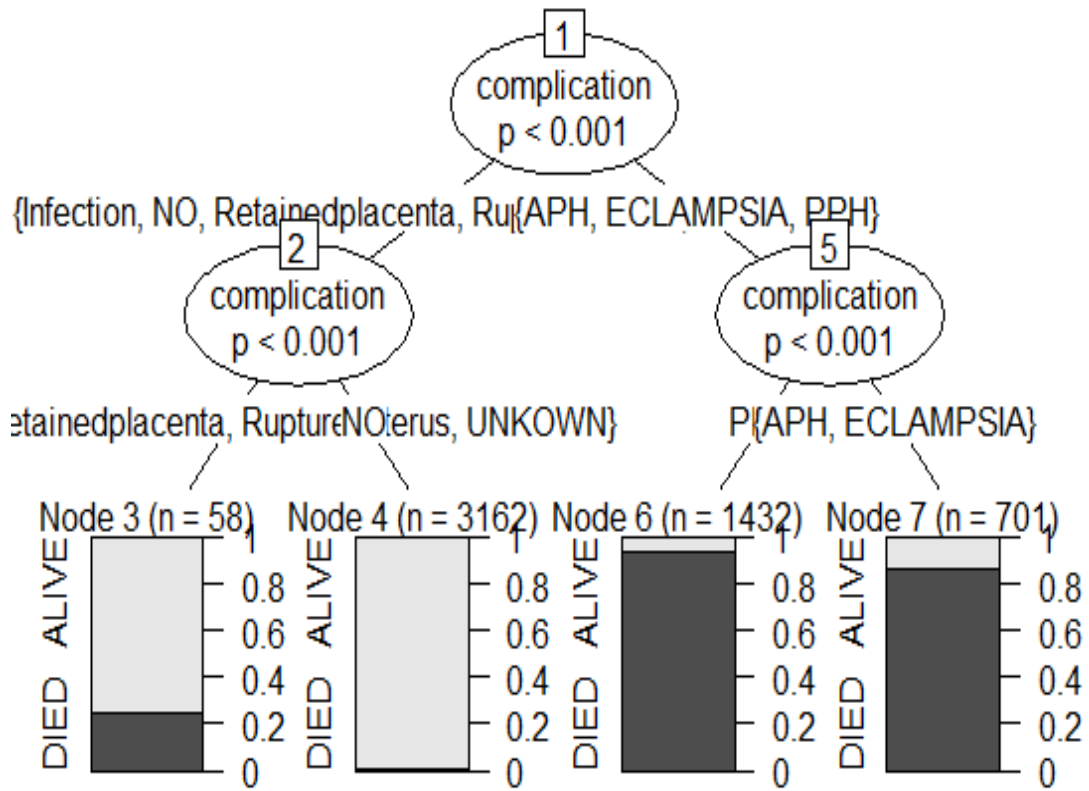


Figure 4.10: Predictive relationship between maternal status and obstetric complication

As observed from Figure 4.10, obstetric complications like PPH, APH, and ECLAMPSIA leads to high maternal mortality.

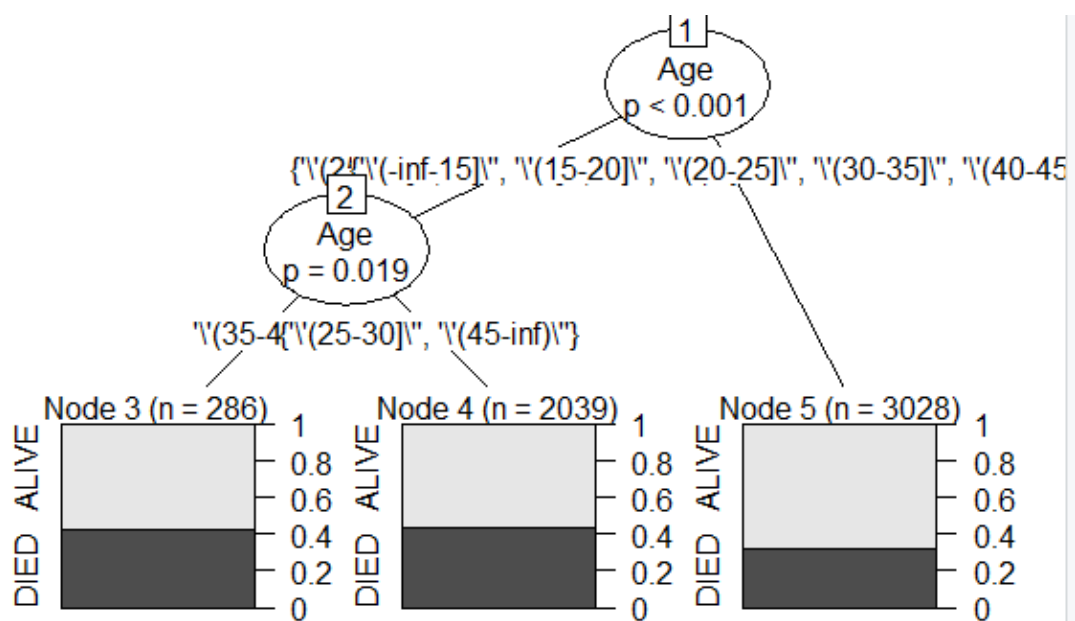


Figure 4.11: Predictive relationship between maternal status and maternal age

Maternal age has a significant relation with maternal outcome or status. As observed from Figure 4.11 above at the age between 25-30 maternal mortality is high. Because women at this age are highly productive than the other age groups. The maternal mortality at age less than 15 and greater than 45 is less as compared with the other age groups.

## CHAPTER FIVE

### 5. Conclusion and Recommendation

#### 5.1. Conclusion

In Ethiopia maternal mortality due to complications during pregnancy and childbirth is high. This study attempted to develop a predictive model for maternal mortality and its determinants using data mining techniques. The experiments are done using three classification algorithms (Naïve Bayes, Decision tree, and Support vector machine) and RStudio open source software used to develop the models. Knowledge discovery in the dataset was employed after the SMOTE technique and feature selection has been applied on the dataset.

After the models were developed using these three classification algorithms, the performances of the models were evaluated using accuracy, sensitivity, specificity, recall, and precision. 10-fold cross-validation and 70/30% percentage split test were adopted as a test option for random sampling of the training and test data samples. All the models performed well in predicting maternal mortality.

The experimental result shows that J48 pruned decision tree model has higher classification accuracy (97.56%) and the SVM radial base function model has the second-highest accuracy (96.64%) and the naïve Bayes model has the lowest accuracy (94.2%). Thus, the researcher concludes that the pruned J48 decision tree classifier is the most effective model to predict maternal mortality patterns and its determinants using a clinical dataset, with 97.56% classification accuracy. The model is used to extract rules and simple to understand. Moreover, the extracted rules from the selected models showed that the major complications that have a high impact on maternal mortality are Maternal condition, Age, and obstetric complications were the major determinant factors of maternal mortality.

The study showed that data mining techniques are highly important to predict maternal mortality patterns. The encouraging result obtained from the three models indicates that

data mining is a method that should be considered to support maternal health care prevention and control activities in Ethiopia.

The result of the study can assist physicians to make a more consistent diagnosis of determinant factors that causes maternal mortality.

## 5.2. Recommendation

This research work uncovers the potential applicability of data mining techniques for predicting maternal mortality patterns. Based on the results of this research, the researcher would like to make the following recommendations concerning the possible application of data mining techniques in maternal mortality reduction and taking preventive actions on the risk factors of pregnant women in Ethiopia. Thus, the researcher forwarded the following recommendations for future research direction in maternal mortality reduction strategies.

- The present study has considered a clinical dataset to use data mining technology in maternal mortality prediction. The use of Epidemiological dataset needs greater emphasis on maternal mortality reduction. Thus, future studies need to derive knowledge and pattern from Epidemiological datasets and compare and integrate them with the result obtained by clinical datasets.
- Although all the three classification algorithms provide a promising result, still performance improvement is needed due to the sensitivity of the domain area. Thus, further extensive experiments should be required by using large amounts of datasets and applying other classification techniques.
- In this study, an attempt has been made to assess the applicability of data mining techniques to predict the likelihood of maternal mortality by using some number of variables that were considered important by domain experts. For many other variables, it is important to build models to obtain better accuracy and performance.
- In this study, attempts were made to explore data mining technique to build maternal mortality predictive modeling based on predefined classes. To use the discovered knowledge, it is important to develop an operational prototype for domain experts.



## REFERENCES

- Banjari, Ines, Daniela Kenjerić, Krešimir Šolić, and Milena L. Mandić. 2015. "Cluster Analysis as a Prediction Tool for Pregnancy Outcomes." 247–52.
- Berhan, Yifru and Asres Berhan. 2014. "Causes of Maternal Mortality in Ethiopia: A Significant Decline in Abortion Related Death." *Ethiopian Journal of Health Sciences* 24(8):15–28.
- Berry, Michael J. A. and Gordon S. Linoff. 2004. "Michael J. A. Berry and Gordon S. Linoff Data Mining Techniques 2nd Edition, Wiley, 2004, Chapter 1."
- Boerma, Ties. n.d. "Maternal Mortality." 4.
- Chawla, Nitesh V, Aleksandar Lazarevic, Lawrence O. Hall, and Kevin Bowyer. n.d. "SMOTEBoost : Improving Prediction of the Minority Class in Boosting."
- Cheng, Jiechao. 2017. "Data-Mining Research in Education Data-Mining Research in Education." (March).
- Cios, Krzysztof J. and Lukasz A. Kurgan. n.d. "1 . Trends in Data Mining and Knowledge Discovery." (Dm):1–26.
- Danilo, Croce. 2010. "Decision Tree Algorithm Weka Tutorial Machine Learning : Brief Summary."
- Du, Jun. n.d. "Data Preprocessing."
- Durairaj, M. and V. Ranjani. 2013. "Data Mining Applications In Healthcare Sector : A Study." 2(10).
- El-hasnony, Ibrahim M., Hazem M. El Bakry, and Ahmed A. Saleh. n.d. "Data Mining Techniques for Medical Applications : A Survey Association : Classification : Clustering : " 205–12.
- Fayyad, Usama, Gregory Piatetsky-shapiro, and Padhraic Smyth. 1996. "From Data Mining to Knowledge Discovery In." 37–54.
- Gebremedhin, Samson. 2018. "Development of a New Model for Estimating Maternal Mortality Ratio at National and Sub- National Levels and Its Application for Describing Sub-National Variations of Maternal Death in Ethiopia." 013:1–18.
- Guyon, Isabelle. 2003. "An Introduction to Variable and Feature Selection 1 Introduction." 3:1157–82.
- Hailemariam, Tesfahun, Million Meshesha, and Alemayehu Worku. 2015. "Health & Medical Informatics Application of Data Mining Techniques to Predict Adult Mortality : The Case of Butajira Rural Health Program , Butajira , Ethiopia." 6(4).

- Han, Jiawei and Micheline Kamber. 2003. "Why Data Preprocessing ?"
- Han, Jiawei and Micheline Kamber. 2006. "Data Mining : Concepts and Techniques."
- Hickey, Stephanie J. 2013. "Naive Bayes Classification of Public Health Data with Greedy Feature Selection." 13(2).
- Hill, Chapel and North Carolina. 2001. "Measuring Maternal Mortality from a Census : Guidelines for Potential Users." (July).
- Idowu, Peter Adebayo. 2018. "Development of a Predictive Model for Maternal Mortality in Nigeria Using." (November 2017).
- Keleş, Mümine Kaya. 2017. "AN OVERVIEW : THE IMPACT OF DATA MINING APPLICATIONS ON VARIOUS SECTORS." 6168:128–32.
- Keller, Frank. n.d. "Naive Bayes Classifiers." 1–22.
- Kim, Hyunjoong and Wei-yin Loh. 2014. *Classification Trees With Unbiased Multiway Splits*.
- Kvåle, Gunnar, Bjørg Evjen Olsen, and Sven Gudmund Hinderaker. 2015. "Maternal Deaths in Developing Countries : A Preventable Tragedy." 15(2):141–49.
- Langleyflamingostanfordedu, P. A. T. Langley, Stephanie Sage, S. A. G. E. Flamingo, and Stanford Edu. 1993. "Institute for the Study of Learning and Expertise 2451 High Street, Palo Alto, CA 94301." (1990):399–406.
- Leung, K. Ming. n.d. "Naive Bayesian Classifier."
- Li, Jundong, Kewei Cheng, Suhang Wang, and Fred Morstatter. 2016. "Feature Selection : A Data Perspective Feature Selection : A Data Perspective." (September).
- Luan, Jing. n.d. "Data Mining Applications in Higher Education."
- Mehta, Rutvij and Nikita Bhatt. 2016. "A Survey on Data Mining Technologies for Decision Support System of Maternal Care Domain." 138(10):20–24.
- Mekonnen, Wubegzier, Damen Hailemariam, and Alem Gebremariam. 2016. "Original Article Causes of Maternal Death in Ethiopia between 1990 and 2016 : Systematic Review with Meta-Analysis." (5).
- Meyer, David. 2019. "Support Vector Machines." 1:1–8.
- Of, Assessment, Maternal Death, Factors Affecting, Maternal Death Surveillance, and I. N. Dire Dawa. 2015. "ADDIS ABABA UNIVERSITY COLLEGE OF HEALTH SCIENCE SCHOOL OF PUBLIC HEALTH ASSESSMENT OF MATERNAL DEATH AND FACTORS AFFECTING MATERNAL DEATH SURVEILLANCE AND RESPONSE SYSTEM A THESIS SUBMITTED TO

THE SCHOOL OF GRADUATE STUDIES OF.”

- Oreski, Dijana and Tomislav Novosel. 2014. “Comparison of Feature Selection Techniques in Knowledge Discovery Process.” 3(4):285–90.
- Parali, Jan and Peter Bednar. n.d. “A TOOL FOR SUPPORT OF THE KDD PROCESS.” (977091):15–27.
- Patel, Nikita. 2012. “Study of Various Decision Tree Pruning Methods with Their Empirical Comparison in WEKA.” 60(12):20–25.
- Pradhan, Manaswini. 2014. “Data Mining & Health Care : Techniques of Application.” 7445–55.
- Press, Dove. 2017. “Trends and Causes of Maternal Mortality in Jimma University Specialized Hospital , Southwest Ethiopia : A Matched Case – Control Study.” 307–13.
- Priyadharsini, C. and Antony Selvadoss Thanamani. 2014. “An Overview of Knowledge Discovery Database and Data Mining Techniques.” 2(1):1571–78.
- Pushpan, Arun and Ali Akbar N. 2017. “Data Mining Applications in Healthcare National Conference On Discrete Mathematics & Computing ( NCDMC-2017 ).” 4–7.
- Ramageri, M. n.d. “DATA MINING TECHNIQUES AND APPLICATIONS.” 1(4):301–5.
- Rani, B. Kavihta and A. Govrdhan. 2010. “Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks.” 02(02):250–55.
- Rokach, Lior and Oded Maimon. n.d. “Chapter 9.”
- Ryan, Shevaun. 2006. “Practical Data Mining Tutorial 4 : Preprocessing.”
- Sahle, Geletaw. 2016. “Ethiopic Maternal Care Data Mining : Discovering the Factors That Affect Postnatal Care Visit in Ethiopia.” *Health Information Science and Systems* (May).
- Setiono, Rudy. n.d. “Feature Selection : An Ever Evolving Frontier in Data Mining.” 4–13.
- Sharma, Himani and Sunil Kumar. 2016. “A Survey on Decision Tree Algorithms of Classification in Data Mining.” 5(4):2094–97.
- Silwattananusarn, Tipawan and Assoc Prof Kulthidatuamsuk. 2012. “Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 To.” 2(5):13–24.
- Thorat, Surabhi and Seema Kute. 2014. “Medical Data Mining Life Cycle and Its

Role in Medical Domain.” 5(4):5751–55.

Wirth, Rüdiger. n.d. “CRISP-DM : Towards a Standard Process Model for Data Mining.” (24959).

Witten, Ian H. n.d. “Data Mining with Weka (Class 1) - 2013.”

Za, Osmar R., South Pole, and Computing Science. 1999. “Chapter I : Introduction to Data Mining.” 1–15.

## APPENDEXES

**Appendix 1:** Sample code used to split the dataset as training and testing set using percentage split and 10-fold cross validation respectively

```
set.seed(1234)
spl<-sample.split(LBSB$mstatus,splitRatio = 0.7)
train=subset(LBSB,spl==TRUE)
test<-subset(LBSB,spl==FALSE)

traincontrol<-trainControl(method = "cv",number = 10)
model<-train(mstatus~.,data = train,method=" ",trcontrol=traincontro
l)
```

**Appendix 2:** selected attributes using gain ratio

```
gainratio<- GainRatioAttributeEval(mstatus ~ Age + mod +  
mcondition + htr + complication,data = train)
```

```
      attr_importance  
Age          0.08076433  
mod          0.05011313  
mcondition   0.52414244  
htr          0.07918931  
complication 0.49938191
```

**Appendix 3:** Sample code used for model development and prediction using pruned J48 decision tree classifier

```
resultJ48 <- Rweka::J48(mstatus~., train, control = weka_control(M = 2, c = 0.5))
```

```
> p<-predict(resultJ48c0.5,test,type = "class")  
> confusionMatrix(table(test$mstatus,p))
```

```
      p  
      ALIVE DIED  
ALIVE 1403   10  
DIED   46  835  
  
      Accuracy : 0.9756  
      95% CI   : (0.9684, 0.9815)  
No Information Rate : 0.616  
P-Value [Acc > NIR] : < 2.2e-16  
  
      Kappa : 0.948  
Mcnemar's Test P-value : 2.91e-06  
  
      Sensitivity : 0.9929  
      Specificity : 0.9478  
      Pos Pred Value : 0.9683  
      Neg Pred Value : 0.9882  
      Prevalence : 0.6160  
      Detection Rate : 0.6116  
      Detection Prevalence : 0.6316  
      Balanced Accuracy : 0.9704  
  
      'Positive' Class : ALIVE
```

**Appendix 4:** Sample of the decision tree generated with 70/30 percentage split Technique.

J48 pruned tree

```

-----
mcondition = STABLE
  complication = APH
    mod = CS: ALIVE (4.0/1.0)
    mod = FORCEPS: ALIVE (3.0/1.0)
    mod = OTHER: ALIVE (1.0)
    mod = SVD
      Age = '\'(-inf-15]\'': DIED (0.0)
      Age = '\'(15-20]\'': DIED (20.0)
      Age = '\'(20-25]\'': DIED (62.0/2.0)
      Age = '\'(25-30]\'': ALIVE (3.0)
      Age = '\'(30-35]\'': DIED (24.0/2.0)
      Age = '\'(35-40]\'': DIED (8.0/1.0)
      Age = '\'(40-45]\'': DIED (0.0)
      Age = '\'(45-inf)\'': DIED (0.0)
    complication = ECLAMPSIA
      Age = '\'(-inf-15]\'': DIED (0.0)
      Age = '\'(15-20]\'': ALIVE (2.0/1.0)
      Age = '\'(20-25]\'': ALIVE (8.0/1.0)
      Age = '\'(25-30]\'':
        mod = CS: ALIVE (1.0)
        mod = FORCEPS: ALIVE (3.0)
        mod = OTHER: DIED (0.0)
        mod = SVD: DIED (21.0/4.0)
      Age = '\'(30-35]\'': DIED (56.0)
      Age = '\'(35-40]\'': DIED (0.0)
      Age = '\'(40-45]\'': DIED (0.0)
      Age = '\'(45-inf)\'': DIED (0.0)
    complication = Infection: ALIVE (6.0)
    complication = NO: ALIVE (3124.0)
    complication = PPH
      Age = '\'(-inf-15]\'': DIED (0.0)
      Age = '\'(15-20]\'': DIED (1.0)
      Age = '\'(20-25]\'': ALIVE (6.0)
      Age = '\'(25-30]\'': DIED (156.0/18.0)
      Age = '\'(30-35]\'': DIED (46.0/1.0)
      Age = '\'(35-40]\'': ALIVE (1.0)
      Age = '\'(40-45]\'': DIED (0.0)
      Age = '\'(45-inf)\'': DIED (0.0)
    complication = Retainedplacenta: ALIVE (16.0/4.0)
    complication = Ruptureduterus: ALIVE (8.0)
    complication = UNKOWN: ALIVE (9.0/1.0)
  mcondition = UNSTABLE
    complication = APH
      mod = CS
        Age = '\'(-inf-15]\'': ALIVE (1.0)
        Age = '\'(15-20]\'': ALIVE (0.0)
        Age = '\'(20-25]\'': ALIVE (12.0)
        Age = '\'(25-30]\'': DIED (1.0)
        Age = '\'(30-35]\'': ALIVE (2.0/1.0)
        Age = '\'(35-40]\'': DIED (12.0)
        Age = '\'(40-45]\'': ALIVE (0.0)
        Age = '\'(45-inf)\'': ALIVE (0.0)
      mod = FORCEPS: ALIVE (2.0)
      mod = OTHER: DIED (166.0)
      mod = SVD
        Age = '\'(-inf-15]\'': DIED (0.0)
        Age = '\'(15-20]\'': ALIVE (8.0)
        Age = '\'(20-25]\'': DIED (17.0)
        Age = '\'(25-30]\'': ALIVE (2.0)
        Age = '\'(30-35]\'': DIED (45.0/12.0)
        Age = '\'(35-40]\'': DIED (11.0/1.0)

```



```

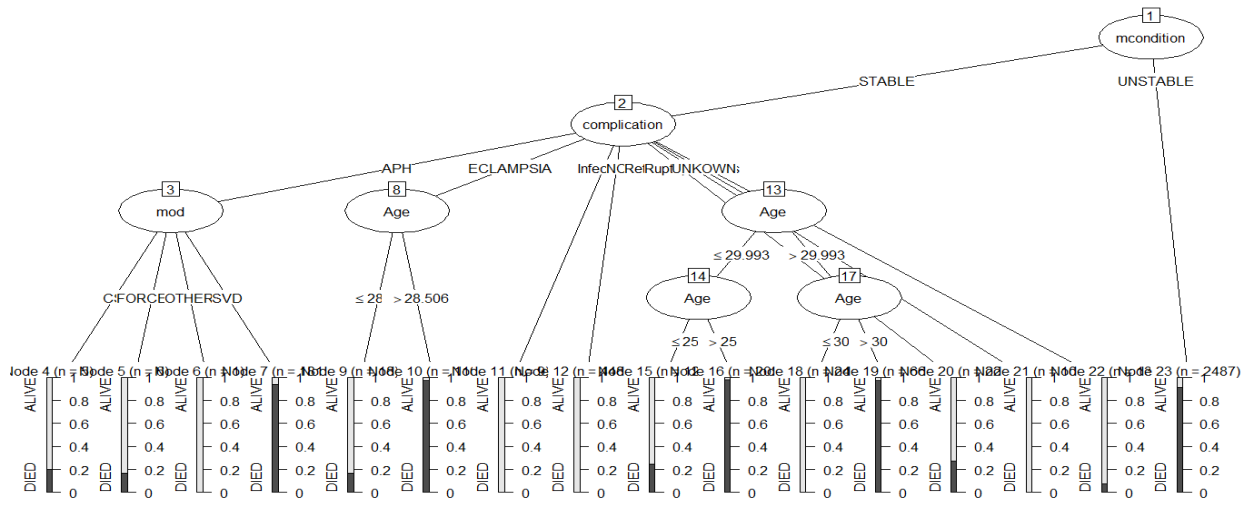
|      |      |      Age = '\'(40-45]\'' : DIED (0.0)
|      |      |      Age = '\'(45-inf)\'' : DIED (0.0)
|      |      |      complication = ECLAMPSIA
|      |      |      |      mod = CS: DIED (115.0/1.0)
|      |      |      |      mod = FORCEPS: ALIVE (12.0/3.0)
|      |      |      |      mod = OTHER: DIED (20.0/3.0)
|      |      |      |      mod = SVD: DIED (59.0/11.0)
|      |      |      |      complication = Infection
|      |      |      |      |      htr = NR: DIED (3.0/1.0)
|      |      |      |      |      htr = R: ALIVE (2.0)
|      |      |      |      |      complication = NO: ALIVE (38.0/1.0)
|      |      |      |      |      complication = PPH
|      |      |      |      |      |      Age = '\'(-inf-15]\'' : DIED (5.0)
|      |      |      |      |      |      Age = '\'(15-20]\'' : DIED (34.0/16.0)
|      |      |      |      |      |      Age = '\'(20-25]\''
|      |      |      |      |      |      |      mod = CS: DIED (321.0/2.0)
|      |      |      |      |      |      |      mod = FORCEPS: ALIVE (2.0/1.0)
|      |      |      |      |      |      |      mod = OTHER: ALIVE (13.0/2.0)
|      |      |      |      |      |      |      mod = SVD: DIED (83.0/11.0)
|      |      |      |      |      |      |      Age = '\'(25-30]\'' : DIED (673.0/16.0)
|      |      |      |      |      |      |      Age = '\'(30-35]\'' : DIED (9.0/1.0)
|      |      |      |      |      |      |      Age = '\'(35-40]\'' : DIED (81.0)
|      |      |      |      |      |      |      Age = '\'(40-45]\'' : DIED (0.0)
|      |      |      |      |      |      |      Age = '\'(45-inf)\'' : DIED (1.0)
|      |      |      |      |      |      |      complication = Retainedplacenta
|      |      |      |      |      |      |      |      mod = CS: ALIVE (2.0)
|      |      |      |      |      |      |      |      mod = FORCEPS: DIED (0.0)
|      |      |      |      |      |      |      |      mod = OTHER: DIED (0.0)
|      |      |      |      |      |      |      |      mod = SVD: DIED (3.0)
|      |      |      |      |      |      |      |      complication = Ruptureduterus: ALIVE (8.0/3.0)
|      |      |      |      |      |      |      |      complication = UNKOWN: DIED (1.0)

```

Number of Leaves : 77

Size of the tree : 93

## Appendix 5: Partial over view of the decision tree



**Appendix 6:** Sample code used for model development and prediction using naïve Bayes classifier

```
nb<-naiveBayes(mstatus~.,data = train)
```

```
p<-predict(nb,test,type = "class")
> confusionMatrix(table(test$mstatus,p))
Confusion Matrix and Statistics
```

	p	
	ALIVE	DIED
ALIVE	1389	60
DIED	73	772

```
Accuracy : 0.942
      95% CI : (0.9317, 0.9512)
No Information Rate : 0.6316
P-Value [Acc > NIR] : <2e-16

      Kappa : 0.875
```

```
McNemar's Test P-Value : 0.2981
```

```
Sensitivity : 0.9586
Specificity : 0.9136
Pos Pred Value : 0.9501
Neg Pred Value : 0.9279
Prevalence : 0.6316
Detection Rate : 0.6055
Detection Prevalence : 0.6373
Balanced Accuracy : 0.9361
```

```
'Positive' Class : ALIVE
```

**Summary(nb)**

Naive Bayes Classifier for Discrete Predictors

```
Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)
```

A-priori probabilities:

Y	ALIVE	DIED
	0.6317953	0.3682047

Conditional probabilities:

Y	Age	'(-inf-15]'	'(15-20]'	'(20-25]'
ALIVE		0.002956830	0.109402720	0.353636901
DIED		0.002536783	0.083206494	0.283105023

Y	Age	'(25-30]'	'(30-35]'	'(35-40]'
ALIVE		0.340626848	0.137196925	0.048492017
DIED		0.446981228	0.112125824	0.061897514

Y	Age	'(40-45]'	'(45-inf)'
ALIVE		0.007687759	0.000000000
DIED		0.007102993	0.003044140

		mod			
		CS	FORCEPS	OTHER	SVD
Y	ALIVE	0.20549970	0.12773507	0.05322295	0.61354228
	DIED	0.38001015	0.02181634	0.20497210	0.39320142

		mcondition	
		STABLE	UNSTABLE
Y	ALIVE	0.95032525	0.04967475
	DIED	0.19025875	0.80974125

		htr	
		NR	R
Y	ALIVE	0.98492017	0.01507983
	DIED	0.92694064	0.07305936

		complication			
		APH	ECLAMPSIA	Infection	NO
Y	ALIVE	0.0156712005	0.0118273211	0.0026611473	0.9346540509
	DIED	0.1780821918	0.1303906646	0.0010147133	0.0005073567

		complication			
		PPH	Retainedplacenta	Ruptureduterus	UNKOWN
Y	ALIVE	0.0248373743	0.0041395624	0.0038438794	0.0023654642
	DIED	0.6839167935	0.0035514967	0.0015220700	0.001014713

## Appendix 7: Sample code used for model development and prediction using SVM

classifier

```
svm<-svm(mstatus~.,data = train)
p<-predict(svm,test,type = "class")
confusionMatrix(table(test$mstatus,p))
```

Confusion Matrix and Statistics

	p	
	ALIVE	DIED
ALIVE	1372	77
DIED	0	845

Accuracy : 0.9664  
95% CI : (0.9582, 0.9734)  
No Information Rate : 0.6316  
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.9292

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9469  
Specificity : 1.0000  
Pos Pred Value : 1.0000  
Neg Pred Value : 0.9165  
Prevalence : 0.6316  
Detection Rate : 0.5981  
Detection Prevalence : 0.5981  
Balanced Accuracy : 0.9734

'Positive' Class : ALIVE

Summary(svm)

Call:  
svm(formula = mstatus ~ ., data = train)

Parameters:  
SVM-Type: C-classification  
SVM-Kernel: radial  
cost: 1  
gamma: 0.05

Number of Support Vectors: 516