Identifying Influential Factors Affecting Pregnancy Events Using Machine Learning Approaches

By

Aymen Tasneem

2016-NUST-MSIT-170713

Supervisor

Dr. Sharifullah Khan

A thesis submitted in conformity with the requirements for
the degree of *Master of Science* in
Information Technology

Department of Computer Science
School of Electrical Engineering and Computer Sciences (SEECS)
National University of Sciences and Technology (NUST)
Islamabad, Pakistan

Certificate of Originality

I, *Aymen Tasneem* hereby declare that this thesis titled “Identifying Influential Factors for Pregnancy Events using Machine Learning Approaches” is my own work and to the best of my knowledge it contains no materials previously published or written by another person , nor material which to a substantial extent has been accepted for the award of my degree or diploma at SEECS, NUST or at any other educational institute, except where due acknowledgment has been made in the thesis. Any contribution made to the research by others, with whom I have worked at SEECS NUST or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project’s design and conception or in style, presentation and linguistics which has been acknowledged.

Aymen Tasneem
2016-NUST-MSIT-170713

This thesis is dedicated to *my beloved parents*

Abstract

Despite of health awareness campaigns and improvements in education system, neonatal mortality is still a critical issue around the world. Out of 140 million children born annually, 4 million die in the first month of their life. This has also become a severe issue in Pakistan with neonatal rate of 44 per 1000 live births. Pakistan is a country where cousin marriage rate is above 60% and is located in such a region where gender preference is common. Studies have suggested that cousin marriage also impacts the pregnancy events. Hence it is a dire need to find out the causes of high neonatal rates and impact of cousin marriage on pregnancy events in Pakistan. Different researches have been carried out on these issues before. Some researchers have investigated the relationship between cousin marriage and adverse pregnancy outcomes while some explored the determinants of child mortality in Pakistan. While these researches focused on specific pregnancy factors such as birth interval and still births, they ignored other important factors like cousin marriage and preterm birth. Some studies have used data with missing factors, such as birth interval, cousin marriage or gestation period; while other studies have mostly applied bivariate or multivariate regression analysis. These techniques have limitations in terms of dealing with categorical data or data with multiple levels of factors. To resolve short comings of the existing studies, we are proposing a framework that will apply association rules, bayesian network and hidden markov model to find associations among different factors in the Pakistan Institute of Medical Sciences (PIMS) hospital dataset. The objectives of this research are (i) to study the effects of different factors that cause neonatal mortality, (ii) cousin marriage impact on neonatal mortality. Finally (iii) to analyze the impact of cousin marriage on gender determination. Data was preprocessed using imputations and models were applied. In order to identify the factors of neonatal mortality, bayesian network and association rules were applied. Bayesian network (BN) produced an accuracy of 94%. Association rules were applied using 'rattle' library and around 9000 rules were generated but only few hold valuable information, such as chances of caesarean delivery are high for short birth intervals and short birth intervals trend has been observed in first pregnancy. To see cousin marriage impact on gender determination, data was distributed on the basis of cousin marriage and non-cousin marriage and was converted into sequential data. Hidden markov model was then applied on each dataset and a comparison was performed to see the impact of cousin marriage on gender identification. Its results indicate that mode of delivery, preterm birth, gestation period and birth interval are the major factors influencing the neonatal mortality. Cousin marriage of couple's parents , place of residence and mother's education are secondary influencing factors. The results also suggest that short birth interval is observed in first pregnancy. Moreover this research also claims that cousin marriage does not play any significant role in the determination of gender. The results will help in improving decision making and making better policies for mother care in order to reduce neonatal mortality rates.

Keywords: *Bayesian network, Maternity dataset, neonatal mortality, Hidden Markov model, Association rules, Cousin marriage, PIMS*

Acknowledgments

I am grateful to Allah who granted me success and it was impossible without His blessings. I would like to give special thanks to my supervisor Dr. Sharifullah Khan and my GEC members, specially Dr. Amanullah Yasin, for guiding me throughout this work. I want to extend my gratitude to all my lovely friends and family for supporting throughout the degree. I would like to pay my thanks to everyone who directly or indirectly contributed in this journey.

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

BN	Bayesian Network
HMM	Hidden Markov Model
PDHS	Pakistan Demographic Health Survey
PIMS	Pakistan Institute of Medical Science
OR	Odd Ratio
WHO	World Health Organization
UN	United Nations

Chapter 1

Introduction

Pakistan has the third highest number of neonatal mortality in the world [1]. To find out the causes behind these high numbers, this research is introducing a new framework. In this chapter, a brief introduction of the issue, motivation behind the selection of this issue, problem statement and objectives of this research work has been discussed. Moreover the framework proposed to fill the research gap has also been discussed.

1.1 Background and Motivation

Neonatal mortality rate is a major concern all around the world. World Health Organization (WHO) and the United Nation's (UN) population division estimate that globally 6.3 million children died in 2017, with newborns accounting for half of these deaths [2]. Every year 2.6 million newborns do not survive their first month of life [3]. Most children under five years of age die due to preventable or treatable causes, such as complications during birth, pneumonia, diarrhea, neonatal sepsis and malaria [3]. The majority of new born deaths, however can be prevented if direct and indirect causes of neonatal deaths can be found [3].

Japan, South Korea and Singapore are among those countries affected by low birth rate caused due to one-child policy by the government [4]. Parents prefer to have son over daughter [4]. When parents find out that their un-born child is a girl through ultrasound, they abort the pregnancy. According to a report [5], many countries are facing "severely skewed" gender ratios. This is a serious issue in Pakistan as well as female neonatal mortality rate is higher as compared to male neonatal mortality rate [6]. One of the emotional trauma that an expecting women face in Pakistan is the gender of the yet to be born child [7]. Gudex [8] suggested that preference is affected by obstetric and socio-demographic factors in women. Nowadays parents tend to do ultrasound during the pregnancy. It is now possible to identify the gender of the unborn child through ultrasound from 18 weeks into pregnancy [9]. In the research carried out by Dow University Hospital, Karachi [9], it was found that 31.4% of the women wanted to know the gender of her unborn child and 15.2% of them wanted to have a son. Therefore the study suggests that half of the women who go for ultrasound preferred to have a son. Moreover Pakistan is among those countries where cousin marriage rate is high [10]. Cousin marriage results in congenital diseases, thalassaemia and other disabilities in the children [11]. Neonatal mortality and haemoglobinopathies are commonly found in the children of such marriages [12].

There are many state-of-the-art techniques [[13], [14], [15],[16] and [17]] that have addressed these issues. In a country like Pakistan where cousin marriage rate is high, gender preference is common and neonatal mortality rate is rising, it is a need of the time to explore these factors in detail. The techniques applied in existing researches are mostly p-values, chi-square, logistic regression models and cox-proportional hazard model. These techniques do not consider casual relationship among different variables (see section 3.2 of Chapter 3), involved in pregnancy events. To find causal relationship among pregnancy variables, bayesian network and association rules will be applied to Pakistan Institute of Medical Sciences (PIMS) dataset. Moreover Hidden Markov Model (HMM) will be applied to find out the impact of cousin marriages on determining gender.

1.2 Problem Statement

Pakistan is among the list of countries where neonatal mortality rate is very high [1]. In order to know what things should be avoided and what steps should be taken, a thorough research on the issues is a dire need of the time. Existing researches mostly used statistical techniques which are unable to give a casual relationship among pregnancy factors. This research focuses on finding causal relationship among pregnancy variables mainly neonatal mortality, cousin marriages and gender determination, by using probabilistic models.

1.3 Research Objectives

This research has three main objectives. The first objective is to identify and discuss the factors playing a role in neonatal mortality in Pakistan. While previous researches carried out on this issue lacks in some way or the other (see section 3.2 of Chapter 3), this research will cover the gap of the previous studies. The second objective is to highlight the impact of cousin marriage factor in neonatal mortality. One of the main reasons for exploring this factor is due to highly prevalence of cousin marriage in Pakistani society (60.5%) [10]. Finally the third objective is to explore how cousin marriage effects gender determination in Pakistani population. They are briefly stated as follows:

1. **Predict factors that cause neonatal mortality:**To find the intervention of a factor and affects of factors on each other
2. **Cousin Marriage impact on neonatal mortality:** To explore cousin marriage impact on neonatal mortality
3. **Cousin Marriage impact on identifying gender:** To analyze the impact of cousin marriage on gender determination

1.4 Proposed Framework

Expecting women now want to know about every event of their pregnancy for a healthy baby. By crunching numbers, data scientists are finding ways to predict pregnancy behavior and better understand complex pregnancy situations. The use of machine learning and decision support technology has given a new dimension to expand the reach of medical professionals in order to improve maternity related decisions. This research proposed a framework that applies bayesian network, hidden markov model and association rules. With the help of these models, applied on maternity dataset, linkages between pregnant woman data and its outcomes can be observed.

1.4.1 Bayesian Network and Maternity Dataset

In order to find out the factors associating with neonatal mortality bayesian network will be used. One of the reasons for using the model is the uncertainty of medical data [18].

When dealing with uncertainty, bayesian network is the right approach [19]. Bayesian network is a model that helps to represent and works through daily life problems, i.e, smoking causes cancer, high cholesterol causes heart attack and other such daily life problems. Beata Reiz [20] used bayesian network to find out causal relationship between different variables in a medical dataset. He predicted surgery survival using this model. The findings proved that bayesian network gives more accuracy as compared to other methods such as logistic regression. Hence the gap in the previous researches that used logistic regression [[13], [14], [15],[16] and [17]] will be filled in terms of accuracy by using bayesian network. We will apply bayesian model to find the causal relationship among different factors involved in neonatal mortality. Moreover cousin marriage influence on neonatal mortality will also be explored using this model.

1.4.2 Hidden Markov Models and Sequence Analysis

One of the objectives of this research is to develop a model that will help in the prediction of the gender of the unborn child. As cousin marriages are prevalent in the country and gender preference is common. Hence the impact of cousin marriage on gender identification needs exploration. To complete this objective, the model used is Hidden markov model (HMM). It is called hidden because the underlying steps between the states are hidden and only the symbols emitted by the model can be observed. The reason for using this model is that it has a strong statistical foundation. Moreover, HMM efficiency in learning algorithms is high. [21].

1.4.3 Association Rules and Variable Dependencies

Apart from hidden markov model and bayesian network model, association rules are also applied on PIMS dataset. It is widely used in medical datasets [22]. By applying it on PIMS dataset the combination of variables mostly occurred together can be known and hence dependency of one variable on another variable can be found [23].

1.5 Organization of Thesis

Chapter 2 contains the background information of bayesian networks, hidden markov model and association rules. Chapter 3 provides the literature review carried out on neonatal mortality and cousin marriage. Chapter 4 describes the proposed framework of this research and workflow of solving the problem. Chapter 5 describes the implementation of the techniques and their results. Chapter 6 will describe the future work and conclusion of this research.

Chapter 2

Background

This chapter describes the basic concepts of the models being used in developing the framework. Three models are used in the proposed framework. The implementation of the models are discussed in chapter 3. The models are Hidden Markov Model, Bayesian Network and Association Rules.

2.1 Hidden Markov Models

It is a model of probability to analyze sequences. It has two states; hidden state and observable state [24]. Model takes transition, emission and initial probabilities as an input and it gives a sequential model as an output. Through the sequence of observations, hidden states can be observed [25]. Figure 2.1 shows the hidden model states [26].

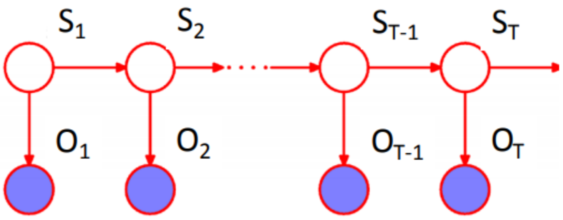
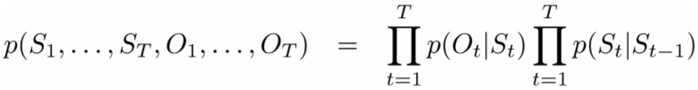


Figure 2.1: Hidden Markov Model States

i. Hidden states: It generates an observable sequence. One hidden state produces one observe state [26]. [S$\hat{t}$]; t = 1,2 ... T . "S" represents hidden state.

ii. Observable states: It is generated from hidden state and depends on all its previous observations.[O$\hat{t}$] ; t = 1,2...T . "O" represents observable state. Figure 2.2 shows observable states, its generation from hidden states and its dependence on previous observations[26].



$$p(S_1, \ldots, S_T, O_1, \ldots, O_T) = \prod_{t=1}^{T} p(O_t|S_t) \prod_{t=1}^{T} p(S_t|S_{t-1})$$

Figure 2.2: Observable States

iii. Transition probabilities: Here pîj is the probability of moving from the hidden state S_1 at time t - 1 to the hidden state S_2 at time t [26]:

$$p(S = j|S^{-1} = i) = p^{ij}$$

We only consider homogeneous HMM, where the transition probabilities are constant over time.

iv. Emission probabilities: It is the probability of the hidden state emitting the observed state: [26]

$$p(O = i|S = r) = q_r^i$$

v. Initial probability: Here πi is the probability of starting from the hidden state "S" [27]:

$$p(S^1 = i) = \pi \ i$$

The Markov assumption of the first order assumes that the transition probability of the hidden state at time t only depends on the previous hidden state at the time point t - 1. The observation at time t depends on the current hidden states irrespective of the previous hidden states and observations [27].

2.2 Bayesian Networks

Bayes or belief network is a joint probability distribution model represented through a directed acyclic graph 'G'. It has a set of random variables. The model represents causal relationship between the variables. The graph comprises of nodes and edges [28]. Nodes (V) represent the variable and edges (E) represent the relationship among the variables. Lack of edges denotes the lack of relation between the variables. In this way, causality is represented in the network. Each node has conditional probability distribution (T), such that G = (V,E) and N = (G,T) where N represents bayesian network [28]. Bayesian network can be created without representing causality, but the representation of causality makes the structure of the Bayesian network more efficient. The objective of the network is to calculate the posterior conditional probability i.e given the evidence which is being observed, what are the unobserved causes P[Cause|Evidence]. Bayesian network works on bayesian theorem.

$$P[Cause|Evidence] = P[Evidence|Cause].P[Cause]/P[Evidence]$$

2.3 Association Rules

Association rules are defined as a set of frequent patterns. Suppose T = [t1, t2, t3, ... tn] is a set of transactions and I = [i1,i2,i3...in] is a set of items. Association rule is defined as A ⇒ B, where A and B are the subset of I and A ∩ B = ∅. A is antecedent and B is consequent of the association rule[29].

Antecedent is the "if" statement of the rule and consequent is the "then" statement of the rule. It is represented as; if item found within the set B then items found in combination with the antecedent i.e set A will be returned. Each item has a measure attached with it that shows its statistical importance which is called "support" [30].Support measures that how frequently the items occur together. Rule is formed when another item, let's say B comes in combination with A and has a statistical measure called "confidence". Confidence of the rule is the conditional probability of the items involved [30]. "Lift" is another measure for the rule and is used to compare the confidence with expected confidence. In other words, lift tells that how better a rule is at predicting the result [29]. Table 2.1 shows the formulas of the statistical measures used in association rules [29].

Criteria	**Formula**
Support	Freq (A , B) / N
Confidence	Freq (A , B) / Freq A
Lift	support / support (A) * support (B)

Table 2.1: Association Rules Measures

Chapter 3

Literature Review

In this chapter, an overview of the existing studies that have been conducted to find out the causes of neonatal mortality, cousin marriages and gender preference are discussed. The existing techniques in finding the causes behind these issues, and their limitations have been critically analyzed. The studies conducted globally and locally (Pakistan) are presented to find out the research gap.

3.1 Pregnancy Events

There are a number of events that are related to healthy pregnancy. Health of the mother before and during pregnancy, financial and educational background of mother and father, pre-natal care, cousin marriage impact and neonatal death are few of the variables that constitutes the pregnancy events. Different researches that were carried out on the pregnancy events mentioned above, will be discussed in this section.

Horsch A. [31] conducted a research to find associations between maternal stress perception and exposure, psychological and physiological stress responses during the pregnancy and pregnancy outcomes. Dataset included 203 expecting mothers of Swiss University Hospital. Results indicate that as far as pregnancy outcomes are concerned strong associations between major life incidents such as depression, anxiety and instrumental delivery were observed. The study concluded that pregnancy event such as maternal stress depends on the hospitalization of baby.

Lakshmi B.[32] studied the complications that occurred during the pregnancy. Decision tree (C5) was used in the research for classification and prediction of risks involved.

They ranked parameters and accessed them which includes; blood pressure, blood glucose level, weight, trimester and month. They compared the results of two datasets: standardized and unstandardized. The unstandardized dataset was obtained from conducting interviews while the standardized dataset was prepared on the parameters suggested by 40 judges who were expert in the field of gynecology. Results showed that decision tree performed better on the standardized dataset as compared to un-standardized dataset. Oppenraaij [33] performed a survey to predict adverse obstetric outcome after early pregnancy events and its complications. They studied the impact of first trimester complications in subsequent pregnancies using data of Cochrane and Medline databases covering the period of 1980 to 2008. The feature set they used includes previous miscarriages, recurrent miscarriages, termination of pregnancies. Results showed that early events in the pregnancy and complication occurred in that time, are the predictors for the adverse pregnancy outcome.

Sable [34] used the data of around 2,378 expecting mothers to examine the relationship between stress, pregnancy attitudes, life events on low birth weight (weight $<$ 1500g). Logistic regression model was applied on the dataset. Results indicate that if mother feels tired most or all of the time during pregnancy, risk of low birth weight gets one and a half times bigger. Unhappiness about the pregnancy, major injury, illness also contribute to low birth weight (LBW). Study concluded that finding an accurate way to predict preterm risk is still a challenging problem. Stephen E.[35] used association rules and data mining to discover interesting patterns in hospital infection control dataset in order to make a new data analysis process.

These studies showed the impact of different variables and the role they play in pregnancy. The studies discussed above shows that during pregnancy an insignificant looking variable can be affecting the pregnancy and thus effecting the health of the baby. The three pregnancy events (variables) that are the focus of this research is neonatal mortality, cousin marriage and gender preference. Section 3.1.1 and 3.2.2 will cover the studies conducted on those pregnancy events that are the focus of this research.

3.1.1 Neonatal mortality

Neonatal mortality is a global issue and researches have been carried out around the world to know the reasons behind this vital issue. Research and exploration to find the causes of neonatal mortality has a significant role in decision making and policy making that helps in the reduction of mortality rates.

Gulam Muhammad [36] used the multilevel logistic regression model to find out the causes of early neonatal mortality in Afghanistan. His study found that multiple gestations, maternal age and birth intervals are the main contributors of early neonatal mortality. The study was limited in terms of biasness of dataset.

Owais A.[13] identified maternal and antenatal factors associated with stillbirths in rural areas of Bangladesh. To determine the causes of death, verbal autopsies were used. A cohort study was conducted and multivariable logistic regression was used to identify the factors. The study found out that asphyxia, sepsis and preterm birth make 35%, 28% and 19% respectively, of the total deaths occurred. The dataset was limited, only 112 cases were examined.

Yu Mu [37] proposed in his research a deep learning algorithm. The algorithm is able to detect and classify adverse pregnancy outcomes. He trained a multi-layer neural network using a dataset of 75542 couples, model outperforms some other algorithms in accuracy but in terms of optimization and interpretability the proposed model was struggling to get better results.

Christiana R. [38] used multilevel logistic regression using a hierarchical approach to find out the determinants of neonatal mortality in Indonesia. They used 2002-2003 demographic and health survey data and found that neonatal mortality rates were higher when mother and father both are employed. Apart from that mothers with history in pregnancy complications, short birth intervals were also the factors for neonatal deaths. Some factors were not examined such as delivery complications, environmental and genetic factors due to limitation of dataset.

Arokiasamy P. [39] analyzed the levels and trends of neonatal mortality in the Indian states. These states constitutes about 60% of infant mortality in India. They studied the impact of bio-demographic and healthcare determinants on neonatal mortality. The dataset they used was ten years old. Cox proportional hazard models were used. The study concludes that antenatal care, delivery assistance and postnatal care are the significant inputs in achieving a reduction in neonatal mortality rates.

Samir B. Kassar [14] identified risk factors for neonatal mortality. They focused on the determinants related to prenatal care, delivery and pregnancy history of mother. The study was conducted in Brazil. Logistic regression was used to analyze the dataset. The sample consisted of 136 cases. The factors identified were hospitalization during pregnancy, poor prenatal care, lack of ultrasound examination and low birth weight. These were the significant factors affecting the mortality rates of neonates. The study has some limitations in terms of data. The impact of the study is claimed to be unsatisfactory and results do not reflect the complexity among some variables.

Jocelyn Finlay [40] analyzed the effects of parity, maternal age and birth interval on the health of child by using the data of Demographic health survey (DHS) from thirty three (33) African countries. Multivariate regression models and Poisson distribution were applied on the dataset. The study found that, in young mothers risk of child mortality is highest for high parity, short birth intervals are negatively correlated with infant mortality. The study was limited in terms of data biasness, exclusion of pregnancies such as miscarriages, stillborn, time duration between subsequent pregnancies.

M. Farrokh [15] focused on whether maternal education and infant mortality have a strong causal relationship or not. Data was taken from Iran demographic and health survey (DHS). The study includes 28 provinces of Iran. Logistic regression was used to measure the effects of these variables. It was seen that infant mortality rates for uneducated mothers is 56 per 1000 live births and for educated mothers are 26 deaths per 1000 live births respectively. The resulting average causal effect was found to be 0.030. This rate means that education of mothers reduces the rate of infant mortality by 30 deaths per 1000 live births. Matching on propensity scores were used to have an estimate about the effect of maternal education on infant mortality. The DHS dataset has not recorded socio economic status and some other variables hence the author used proxies for these variables.

In the research done by Naddav.Y [16], they applied a number of classification techniques (e.g., Naive Bayes, Decision trees, SVM, logistic regression, and associative classifier) on a dataset of historic maternal and newborn records to predict preterm birth. But the classifiers performed poorly on the dataset even after performing feature selection through contrast sets. Hence their work indicates that predicting preterm is a challenging problem.

Sarah Rabbani and Abdul Qayyum [17] investigated the determinants of child mortality in Pakistan. They used data of Pakistan Demographic Health Survey (PDHS) of the year 2006-07. Binary logistic Regression model was used using maximum likelihood method having Bernoulli distribution. The factors that have significant impact on child mortality in Pakistan are wealth index, mothers education, age of parents and exposure to media. Another research that was conducted in Pakistan [41] was related to determinants of neonatal mortality. The study concludes that antenatal care, proper training of health care providers can help in the reduction of mortality rates. With the help of multivariate cox proportional hazard models the study found that wealth index, male infants, first order baby, complications during pregnancy are the main causes of high mortality rates in Pakistan. They were unable to find the effects of environmental causes due to its unavailability in PDHS dataset.

Mosley [42] presented a structure that analyzed factors of child mortality and observed both biological and socioeconomic factors. The emphasis of their research was on individual level decision making. Environmental and geographical factors also play a vital role in determining the health of a child but the research did not include it due to lack of data. Summary of literature review related to neonatal mortality can be found in the Table 3.1

Sr.No	Reference	Dataset	Technique	Limitation	Findings
1	Owais A.(2013)	2011-2012 Bangladesh	Multivariate Logistic Regression	Dataset was limited	Birth Asphyxia, sepsis, preterm birth were the main factors for neonatal mortality.
2	Gulam Muhmmad (2018)	2015 Demographic Health Survey (DHS)	Multilevel Logistic Regression/ Odd Ratios (OR)	Limitation and Biasness of dataset	Multiple gestations, maternal age and birth intervals were the factors effecting neonatal death rates.
3	Yu Mu (2018)	2015 NFPC	Neural Network and Decision Tree	Struggling to get better and interpretable results	Birth Defect, low birth weight, still birth were examined
4	Christiana R (2008)	2003 Demographic Health Survey	Multilevel logistic regression	Some determinants of neonatal mortality were not available in the dataset	Short birth interval is one of the many reasons of neonatal mortality.
5	Arokiasamy P. (2008)	1998-1999 NFHS	Cox Proportional Hazard Models	Dataset used was ten years old	Antenatal and post natal care will help in reducing neonatal deaths

Table 3.1: Summary of literature review for Neonatal mortality

Continuation of Table 3.1

Sr.No	Reference	Dataset	Technique	Limitation	Findings
6	Samir B Kassar et. al (2013)	Dataset set was taken from Brazilian database	Logistic Regression	Results are subjected to recall bais. Variables such as birth interval or cousin marriage were not included in the study	Inadequate prenatal care, low birth weight major factors of neonatal mortality.
7	Jocelyn Finlay (2017)	DHS of 33 African countries	Multivariate Regression Models	Exclusion of pregnancies with miscarriages, stillborn and abortion	In young mothers risk of neonatal mortality is high with high parity
8	M Farrukh (2009)	Iran DHS	Logistic Regression	No information on socio economic status of households	Education of mother reduces infant mortality
9	Naddav Y. (2008)	1992-2003 Northern and Central Alberta Perinatal Outreach Program	SVM, NB, C5, Logistic regression	Prediction performance was not satisfactory, and even doing feature selection by contrast sets does not improve the results.	Predicting preterm is a challenging problem
10	Sarah Rabbani (2015)	2006-2007 PDHS	Binay Logistic regression	PDHS recorded neonatal mortality rates are lower as compared to surveillance dataset.	Mother's education, age, wealth index are factors associated with neonatal mortality

Table 3.2: Summary of literature review for Neonatal mortality (continued)

Continuation of Table 3.1

Sr.No	Reference	Dataset	Technique	Limitation	Findings
11	Nisar B Y. (2014)	2006 - 2007 PDHS	Multivariate cox Proportional hazard models	Environmental factors were not available in PDHS dataset	Wealth index, male infants, first order baby, complications during pregnancy are the main causes of high mortality rates in Pakistan
12	Mosley (2003)	-	Framework (Social and Medical Science)	Only 3 maternal factors were explored such as age, parity and birth interval	Analyzed biological and social determinants of mortality

Table 3.3: Summary of literature review for Neonatal mortality (continued)

3.1.2 Cousin marriage and Gender Preference in Pakistan

The effects of cousin marriage on population health and reproduction are a major issue in South Asian and Middle Eastern countries. In the research done by Dow University Hospital, Karachi, it was found that 31.4% of the women wanted to know the gender of her unborn child and only 15.2% of them wanted to have a son. Therefore the study suggests that half of the women who go for ultrasound preferred to have a son [43]. In Pakistan desire for sons was mainly affected by socioeconomic class, financially struggling mothers have more desire to have sons [44]. Mushfiq [45] study suggests that children born to consanguineous marriages had 5.7% higher rate of genetic illness as compared to those children born to non-consanguineous marriages in Pakistan [46]. Another study suggests that women married to their first cousins have higher chance of miscarriages, neonatal and post-neonatal mortality. Saad [47] explored the relationship between miscarriages and cousin marriage. The analysis was performed on Qatari population. 92 women who were in consanguineous marriage were compared with 92 women who got married outside the family. The obstetrical history of both sampled data was same, women in both the samples suffered three or more early pregnancy loses. The study suggests that there was no difference in medical complications between the two samples. A number of pregnancy losses was similar in both groups.Omer S. [48] study shows negative impact of cousin marriage on various variables such as mortality, low birth rates, high rates of stillbirths and miscarriages. However this study limits itself in terms of data by not examining environmental factors. Another study conducted by Kuntla [49] investigated the relationship between cousin marriage and adverse pregnancy outcomes. Bivariate, trivariate regression models were applied on Indian dataset. Analysis revealed that women married to their cousins have high rates of stillbirth, abortions and miscarriages as compared to those women who are not married to their cousins. Another study came out in the year 2019 [46], it concluded that studies carried out previously on the impact of cousin marriage that fails to add socio-economic were baised and falsely precise and the relationship between parent's cousin marriage and child's health is not as statistically precise as claimed by previous studies. Summary of the related work can be found in Table 3.4

Sr.No	Reference	Dataset	Technique	Limitation	Findings
1	Omer S. and Farooq S (2016)	2012 - 2013 PDHS	Chi-square Test	PDHS recorded NMR are lower as compared to surveillance data. Environmental factors are not included	Consanguinity contributes to the negative effects on maternal and child health.
2	Mushfiq Mubarak (2019)	2009 - 2010 Pakistan 2006-2007 Bangladesh	Two-stage least squares (2SLS)	Data has weakness in terms of recording genetic illness, based only on reports by parents and not a medical diagnosis.	Study did not find statistically significant relationship between consanguineous marriage and negative health outcomes for children
3	Saad et al (2002)	Qatari women -184 cases	Mann Whitney U-test	Sample set was small (92) in terms of population epidemiology	No difference in complication between two samples
4	Kuntla S. (2013)	2005 India Human Development Survey	Bivariate, Trivariate, Cox proportional hazard model	-	Women in cousin marriage may suffer from stillbirths, miscarriages and abortions

Table 3.4: Summary of literature review for Cousin Marriages

3.2 Critical Analysis

The critical analysis includes the analysis and limitations of the techniques used in the studies discussed and limitations of the studies itself.

Jasim A.[50] conducted a research in 2018 and used enhanced surveillance of births and deaths in order to determine accurate maternal and neonatal mortality. As compared to previously collected data, all the mortality rates recorded through this system were higher than the estimates given by Pakistan Demographic Health Survey (PDHS). Neonatal rates were 40 as compared to 20 per 1000 live births. The researches [17], [41] used data of Pakistan Demographic Health Survey (PDHS) 2006-07. Hence it can be said that the results of the previously done researches using the data of PDHS are not reliable and improvement in completeness of data is required to get accurate results.

The studies [13], [36] and [14] used the data with some missing variables such as birth interval, cousin marriage and gestation period. Some studies [42] ignored biological factors, only three maternal factors were examined. Hence the dataset they used was limited in terms of determinants of neonatal mortality. Mushfiq et al. [46] empirically investigated the social and economic causes of cousin marriages. He concludes that the estimates showed by the studies done previously on the impact of cousin marriages on child health were falsely precise. Hence on the basis of this study, we can conclude that the studies conducted previously on the impact of cousin marriage are not reliable and new study in this area needs to be carried out. In order to analyze the impact of cousin marriages without any biasness, approach should be revised and much larger samples are required to determine the causal relationships more precisely. Moreover cousin marriage impact on gender determination also needs to be explored.

Existing studies carried out on the issues so far, have mostly applied bivariate or multivariate regression analysis, some used logistic regression and cox proportional hazard models while others [[13], [36], [14]] are using logistic regression to find the causes of neonatal mortality. Logistic regression only deals with binary data, for multi-level categorical data this model is not appropriate. The study [38] applied multilevel logistic regression; it can give cause and effect relationship between two variables but to find out the cause and effects of multiple variables, this model fails to give the desired results [51]. Multilevel logistic model deals with multiple variables only when those variables do not depend on each other moreover adding more variables will cause over-fitting.

Cox proportional hazard models has also been used to conduct studies related to finding causes of neonatal mortality. The model takes the input variables that are time dependent and assumes that hazard function depends on the values of covariates and baseline hazard. The violation of the assumption can result in false deductions [52].

3.3 Discussion

The critical analysis presented above shows that, it is a need to analyze the factors involved in pregnancy events in detail. The dataset should be without any bias and such techniques should be applied that can analyze the complex relations among different variables. Hence data mining techniques were introduced to analyze the pregnancy events that will be able to cover all the weaknesses discussed in this chapter and this research will provide a tool that will address the issues of causality. This research meets the critical need to update the knowledge and information on pregnancy events and its effects on pregnancy outcomes by analyzing a comprehensive assessment of Pakistan Institute of Medical Sciences (PIMS) dataset. PIMS dataset considers the variations in the data by covering all the states and socioeconomic groups across Pakistan.

Chapter 4

Proposed Framework

In this chapter, the framework adopted to achieve research objectives is discussed. The framework is designed to cover the gap found in existing studies (see chapter 3). In section 4.1, the proposed framework and the steps followed to achieve the objectives have been discussed. Section 4.2 is about data acquisition. Section 4.3 discusses about data preprocessing and explains how data cleaning was performed. The next section explains feature engineering as some variables were engineered to observe their impact on the pregnancy outcomes. In section 4.4, models applied to carried out the research and their implementation have been covered.

4.1 Framework Steps

The research procedure followed to achieve the objectives is discussed in this and coming sections. The proposed framework is divided into different phases shown in Figure 4.1.

- Data Acquisition
- Data Pre-processing
- Feature Engineering
- Feature Selection
- Model Implementation

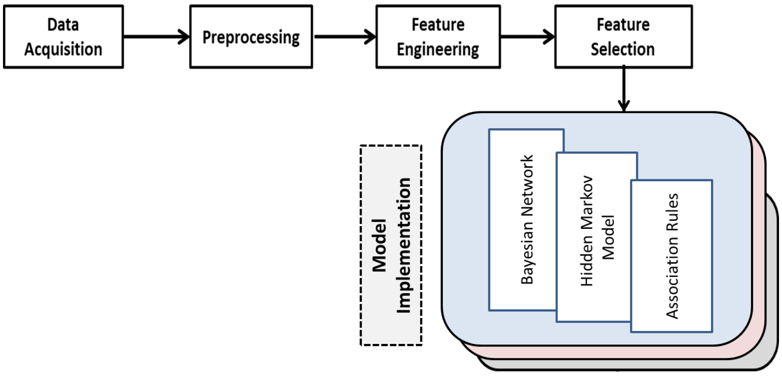


Figure 4.1: Framework used to analyze Pregnancy Events

4.2 Data Acquisition

Data was acquired from Pakistan Institute of Medical Sciences (PIMS) Hospital Islamabad that covers the time period between 1983 to 2014. PIMS hospital is a government hospital whose expenses are low, so patients from all over the country come to this hospital for treatment. Hence, in terms of region, our data covers almost all provinces of Pakistan. The data consists of 5000 pregnant women. Four pair of students collected the data in around 8 years. Every pregnancy history of a single mother was recorded, from her first pregnancy, its outcome, mode of delivery, duration of pregnancy, reasons of the death of the child (in case of the child died after delivery), the year pregnancy occurred, the year the couple got married till her last pregnancy. Basic demographic information was also recorded for all consented mothers. Data also has record of the education, blood group, age of father and mother, occupation of father and mother, family economic status, family type, the province/district where the family resides and whether the marriage was taken place in family or outside the family. Maternal and Neonatal outcomes were recorded at the time the mother gave birth. This all makes up to total 12,593 entries in the dataset. In terms of completeness and socio economic factors, we can claim that our data is comprehensive in contrast to dataset of previous studies.

4.3 Data Preprocessing

Demographic data and pregnancy records were combined into a single analysis dataset using R, with one observation for each birth outcome. Data consists of missing values, invalid values of some features and typos. Two approaches were used to deal with missing values; replacement with NA and imputations. Few values were replaced with "NA" and invalid values were removed using imputations. It was necessary to perform imputation as amount of missing data was significant. Replacing all missing values could have introduced biasness in our resulting model. To implement imputation, R has a package MICE that allowed us to do multiple imputation. This imputation works by filling the missing data multiple times. Multiple imputations are better than single imputation as it deals with uncertainty of missing values in a better way [30]. MICE provides a nice piece of flexibility [53]. We used polytomous regression with multiple imputed datasets. This is because polytomous regression deals with variables having factors two or greater than two. This regression model fits best with PIMS dataset as the dataset was mainly based on variables having multiple level factors. Twenty one variables with 12,593 observations were analyzed.

4.4 Feature Engineering

PIMS dataset has the variable "pregnancy duration", two variables were engineered from this variable e.g " Gestation Period" and "Peterm Birth" as both of the variables play role in effecting pregnancy outcomes. Most of the variables were categorical except for "Gestation Period" and "Pregnancy Number". Continuous variables were discretized using standard categories. Gestation period was divided into three categories first 3 months were categorized as "first trimester", next three months were categorized as "second trimester" and last three months as "third trimester" . The categories for preterm births were used according to World Health Organization (WHO) report (shown in Table 4.1).

In the original dataset, birth spacing between subsequent pregnancies was not recorded. This variable was derived from two other variables i.e the year of marriage of the subject and the year of delivery as mentioned in Table 4.2.

PIMS dataset is in two different excel sheets; one contains the demographic data of all patients and other contains the pregnancy record. To execute the model, data was required to be in one file. The identification factor in both the files for the patient was "Performa number" On the basis of which integration of both the files is performed. For

Table 4.1: Engineered variables Gestation Period and Preterm Birth

Pregnancy Duration	Gestation Period	Preterm Birth
7 Months	2nd Trimester	Very Preterm
9 Months	3rd Trimester	Late Preterm
6.5 Months	2nd Trimester	Extremely Preterm
8.5 Months	3rd Trimester	Moderate Preterm
3 Months	1st Trimester	Extremely Preterm

Table 4.2: Feature Engineering Performed on "Birth Space"

Performa Number	Pregnancy Number	Year of Marriage	Year of Delivery	Birth Space
10	1	2000	2001	1 Year
10	2	2000	2003	2 Year
11	1	2003	2005	2 Year
11	2	2003	2007	2 Year
11	3	2003	2008	1 Year
11	4	2003	2010	2 Year

integration, joins are used [54]. Required file was acquired using left join query.

4.5 Feature Selection

After preprocessing of data and feature engineering, next step was to select the effective features.Therefore, data was collected by considering those features that would impact pregnancy events in direct or indirect way. PIMS dataset consists of only those features that might play some role in affecting pregnancy events. All the features in PIMS dataset were included in the study which makes it to 21 features in total.

4.6 Model Implementation

There are many statistical techniques which were applied previously to analyze pregnancy events. Some studies also used machine learning and data mining techniques but they were only used to analyze mother's health and did not focus on neonatal mortality. In our research, we have applied three different machine learning models to analyze our dataset; Hidden Markov model (HMM), Bayesian network (BN) and Association rules (AR). Libraries were installed in RStudio to implement HMM and AR models. HMM can analyze the sequence of events in an efficient way [24]. It has been widely used in bio informatics and life sequences [25]. HMM is a foundation for making probabilistic models of linear sequence [55]. We have used HMM to analyze the sequence of pregnancies,i.e, 1st pregnancy, 2nd pregnancy, 3rd pregnancy and so on. Moreover, HMM also helped in analyzing the affect of subsequent pregnancies on each other.
GeNIE Modeler has been used to implement Bayesian Network (BN). BN is a graphical model to find causes and effects of relationships and deals well with uncertain data. Our research focuses on finding the causal relationship hidden in PIMS dataset. Although there are many models that are used to represent uncertain domains, such as neural networks, decision trees and markov model but the literature for BN is the only one that represent and learn directed causal relationships among variables [56]. This ability of BN aligns well with our main research goal. The reasons behind choosing BN model for our research are:

- BN is capable of displaying relationships intuitively
- It is bidirectional, hence represent causal relationships effectively
- It represents not only direct causation but indirect causation too
- With the help of established theory of probability, it handles uncertainty

RStudio library "Rattle" was applied to implement AR. Knowledge discovery from data is a process used by many researchers to explore their dataset and find patterns [57]. The most popular data mining tasks are classification, clustering and association rules generation. Association rules are widely used in the field of data mining and can also be used for classification [29]. To implement association rules, apriori algorithm is used to generate the rules.

Chapter 5

Results Evaluation and Validation

To find the factors influencing the pregnancy events, bayesian network, hidden markov model and association rules were applied as discussed in the previous chapter (see chapter 4). With the aim of finding the causal relationship between different variables, predictive models were used in order to observe those factors. In this chapter the implementation and results will be discussed along with the evaluation measure for the results. Accuracy was used to validate the models implemented. After model evaluation, comparison of different models were also discussed.

5.1 Experimental Protocols

Experimental protocols includes software specification and hardware, used for the implementation of models.

- **Software**: Following tools were used to implement the models; R Studio and BayesFusion.

  1. **RStudio**: Evironment (R Desktop) is a statistical tool for data analysis and manipulation. One out of many other reasons for using R is its strong visualization capabilities. It can create visual plots of complex and large datasets that helps in identifying the patterns and anomalies hidden within the data [58]

  2. **BayesFusion**: It is a machine learning based software and provides an artificial intelligence modeling of Bayesian networks. BayesFusion provides GeNIe

modeler to develop graphical decision theoretic models and makes Bayesian networks in an efficient manner.

- **Hardware** The system used to implement these models has the specifications as; 4GB RAM, Intel core i5, 2.20 GHz Processor, 64 bit OS, Windows 10 Pro

5.2 Implementation

5.2.1 Hidden Markov Model Implementation

By using HMM model, we will analyze the impact of cousin marriage factor on the pregnancy outcomes. Our data set was in the form of longitudinal data and was on excel sheets. As we wanted to see the impact of factors on a sequence of pregnancies, hence we used HMM. For this purpose, firstly we need to transform this data in sequential form so that we could run hidden markov model. We have a total number of 5000 gender sequences of pregnancies. As all pregnancies included in our data set are of different sequences hence this model is suitable as it can cater sequences of variable lengths (shown in table 5.1) and can analyze them efficiently. Other than sequence analysis, HMM's are widely used in bioinformatics [26]. R packages such as “TraMiner” and "seqHMM" facilitates with the whole analysis process of this model [25]. The implementation steps are shown in figure 5.1. Firstly, raw data was converted into a meaningful data. Longitudinal data was converted into sequential data. The data for cousin marriage and non-cousin marriage was put in separate sheets to generate separate models for them. Packages were installed and libraries were imported. Transition, emission and initial Probabilities were calculated to built the model. Emission probabilities were calculated manually from data sheet. The model calculated transition probability and initial probability from the data set.

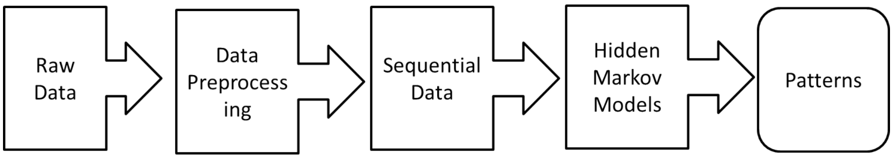


Figure 5.1: Flowchart of implementation of HMM

The observable state is the gender sequences and hidden state is cousin marriage factor. The impact of cousin marriage on gender and pregnancy outcomes were observed. The

Table 5.1: Structure of data for Hidden Markov Model

Sr. No.	1st Pregnancy	2nd Pregnancy	3rd Pregnancy	4th Pregnancy
1	Son	Daughter	Son	X
2	Daughter	Daughter	Daughter	Son
3	Son	Son	X	X

packages used are:

1. **TraMiner**: It is a package in R use for visualizing and analyzing the data with categories. The package helps in discovering knowledge from sequence of events such as life events or DNA sequences. In this research pregnancy events will be examined. The name TraMiner stands for 'life trajectory miner' for R. The reason for using the package was not only to understand the sequences but also in [59]:
    - Handling multiple number of states
    - Displaying the frequent sequences
    - Summaries of sequence states
2. **seqHMM**: This package is designed to efficiently handle sequences using HMM. seqHMM means Sequence Hidden Markov Models. The package provides good graphical options to display the sequences in an understandable way [26]. Functions and methods of seqHMM package used in the analysis are shown in table 5.2:

Table 5.2: TraMiner and seqHMM functions used

Usage	Functions/Methods
Model Construction	build$_{h}mm, build_{m}m$
Model Estimation	fit$_{m}odel$
Model Visualization	plot
Model Inference	Summary, BIC
Transition Probabilities	seqtrate

3. **MICE** : This package helped in dealing with missing data more efficiently. It performs multiple imputations to impute the missing values in the data set. Method uses fully conditional specification where every incomplete variable is imputed separately [60]. The algorithm can impute multiple types of data, such as binary, categorical and continuous data. 'Polytomous logistic regression – polyreg' was used as it works well for un-ordered categorical data.

4. **Dplyr**: It is a package to manipulate the data and transform it. This transformed data is then used to perform functions and methods. It allows fast data exploration and data handling [61]. Two of its functions were used to plot the graphs:

Table 5.3: Dplyr package functions

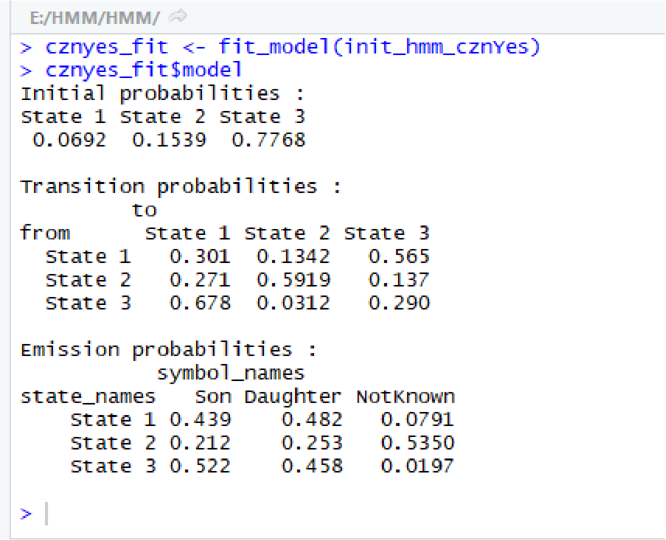
Usage	Functions/Methods
Plotting graphs	ggplot
Joining tables	left$_join$

Table 5.4: Division of data set

Dataset	**Number of Cases**
Cousin Marriage	3054
Non Cousin Marriage	2202

There are a total of 3054 cousin marriage cases and 2202 non-cousin marriage cases (as shown in Table 5.4) in the data set. By using "fit-model" [?] function parameters of the model was learned (as shown in Figure 5.2 and 5.3) and then used the function "build-hmm" to finally build our model on the basis of learned parameters. To plot the resulting model "plot" function was used. This function provides various possibilities to plot graph in an efficient way hence R provides much better visualization of models. The one used to build the model is:

plot(x, layout = "horizontal", vertex.size = 40, vertex.label = "initial.probs", vertex.label.dist = "auto", loops = FALSE, edge.curved = TRUE, edge.label = "auto", edge.width = "auto", edge.arrow.size = 1.5, label.signif = 2, label.scientific = FALSE)



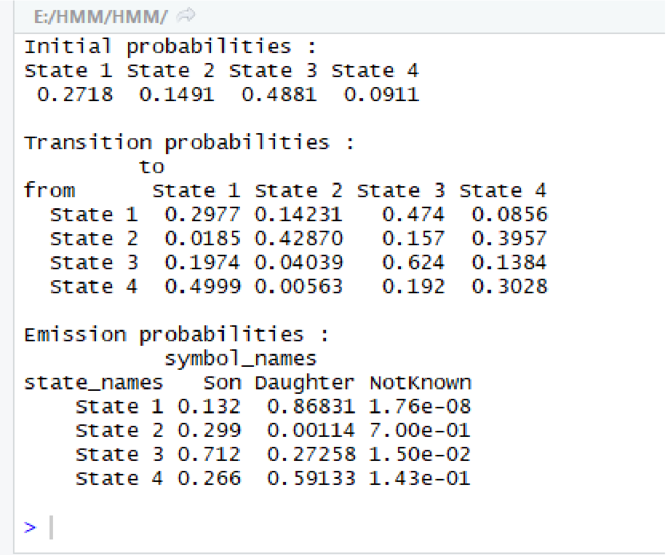
```
> cznyes_fit <- fit_model(init_hmm_cznYes)
> cznyes_fit$model
Initial probabilities :
State 1 State 2 State 3
 0.0692  0.1539  0.7768

Transition probabilities :
         to
from      State 1 State 2 State 3
  State 1   0.301  0.1342   0.565
  State 2   0.271  0.5919   0.137
  State 3   0.678  0.0312   0.290

Emission probabilities :
           symbol_names
state_names   Son Daughter NotKnown
    State 1 0.439    0.482   0.0791
    State 2 0.212    0.253   0.5350
    State 3 0.522    0.458   0.0197

>
```

Figure 5.2: Screenshot of 'fit-model' function for cousin marriages



```
Initial probabilities :
State 1 State 2 State 3 State 4
 0.2718  0.1491  0.4881  0.0911

Transition probabilities :
         to
from      State 1 State 2 State 3 State 4
  State 1  0.2977 0.14231   0.474  0.0856
  State 2  0.0185 0.42870   0.157  0.3957
  State 3  0.1974 0.04039   0.624  0.1384
  State 4  0.4999 0.00563   0.192  0.3028

Emission probabilities :
           symbol_names
state_names   Son Daughter NotKnown
    State 1 0.132  0.86831 1.76e-08
    State 2 0.299  0.00114 7.00e-01
    State 3 0.712  0.27258 1.50e-02
    State 4 0.266  0.59133 1.43e-01

>
```

Figure 5.3: Screenshot of 'fit-model' function for non-cousin marriages

5.2.2 Implementation of BN Model

The bayesian network was learned using different algorithms and then comparison between different algorithms was performed. The algorithms performed were Bayesian Search and PC. With these algorithms, structure was learned. A brief introduction of algortihms and the parameters they require are given below:

a. Bayesian Search:

Bayesian search is one of the most popular algorithms. It uses hill climbing procedure by restarting randomly [62]. It takes the following parameters as input:

- Max parent count
- Iterations
- Sample Size
- Seed
- Link Probability
- Prior Link Probability
- Max Time
- Use Accuracy as Scoring functions

Default values for all parameters were used except for the 'sample size'. It produced good accuracy.

b. PC:

The PC-Algorithm, was initially developed by Spirtes and was implemented in the BayesFusion GeNIe, by Saeed Amizadeh, Steve Birnie, Jeroen J.J [62]. It takes account of the independences observed in data. The PC-Algorithm is a constraint-based algorithm and it calculates the conditional probability distribution on all variables by using a series of conditional independence tests [63]. To extract the patterns direct acyclic graph (DAG) was created. The reason for using this algorithm is that it was experimentally verified by Voortman [64] that this algorithm is fairly robust to the assumption of multi-variate Normality. It has the following parameters:

- Max Adjacency Size
- Significance Level
- Max Time

Estimating the Effects:

Limitation for bayesian network model was missing values and continuous data. The data was converted into different categories and missing values were handled using imputations as discussed in the previous chapter (see section 4.2). After performing pre-processing and model implementation, the required cause and effect graph was created. The estimates of effects of different factors on each other

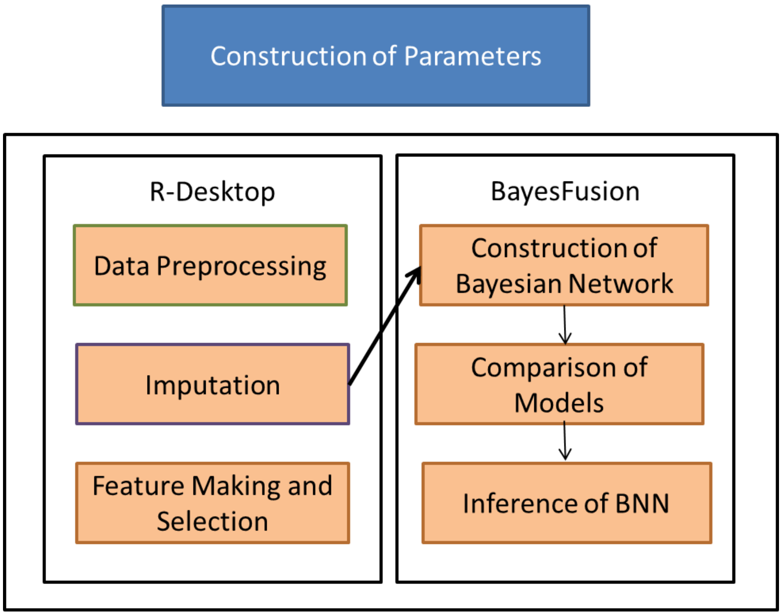

Construction of Parameters
R-Desktop
Data Preprocessing
Imputation
Feature Making and Selection
BayesFusion
Construction of Bayesian Network
Comparison of Models
Inference of BNN

Figure 5.4: Parameter construction for Bayesian network

were calculated and analyzed. Figure 5.4 shows the parameter construction for bayesian network. The steps to create the bayesian network is shown in figure 5.5.

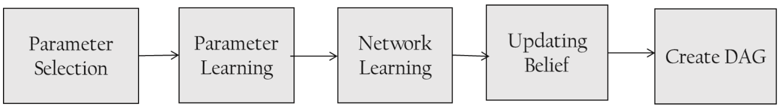
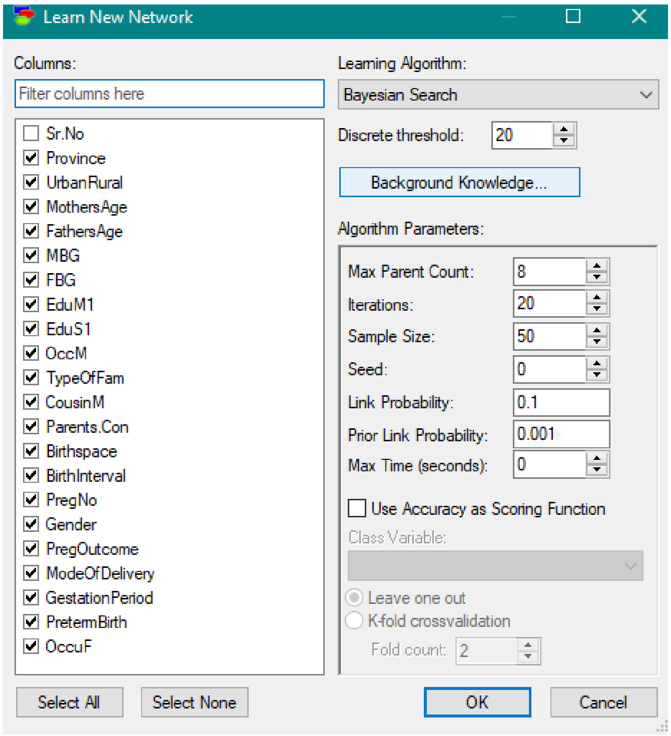


Figure 5.5: Flowchart of Bayesian Belief Network

To find the causal relationship between different factors bayesian networks provides a fairly well developed and extensive structure [65]. Dealing with medical data means nothing can be said with 100% guarantee. Bayesian network handles the data with uncertainty efficiently [66] and is a widely used method for the representation of knowledge [67] hence using Bayesian network for PIMS data set was an appropriate choice. The model was implemented in “BayesFusion - GeNIe Modeler”. The model took the imputed data file as an input. Features/variables were selected for the model was to be built (as shown in Figure 5.6. After this, algorithm to be applied was selected and if background knowledge is required or not. Background knowledge to the model was acquired from Quaid-e-Azam University Professor, Dr. Sajid Hassan; who is a domain expert in this research area (Figure 5.7).



Columns:

Filter columns here

- ☐ Sr.No
- ☑ Province
- ☑ UrbanRural
- ☑ MothersAge
- ☑ FathersAge
- ☑ MBG
- ☑ FBG
- ☑ EduM1
- ☑ EduS1
- ☑ OccM
- ☑ TypeOfFam
- ☑ CousinM
- ☑ Parents.Con
- ☑ Birthspace
- ☑ BirthInterval
- ☑ PregNo
- ☑ Gender
- ☑ PregOutcome
- ☑ ModeOfDelivery
- ☑ GestationPeriod
- ☑ PretermBirth
- ☑ OccuF

Learning Algorithm: Bayesian Search

Discrete threshold: 20

Algorithm Parameters:

Max Parent Count: 8

Iterations: 20

Sample Size: 50

Seed: 0

Link Probability: 0.1

Prior Link Probability: 0.001

Max Time (seconds): 0

☐ Use Accuracy as Scoring Function

Class Variable:

○ Leave one out

○ K-fold crossvalidation

Fold count: 2

Figure 5.6: BayesFusion input file

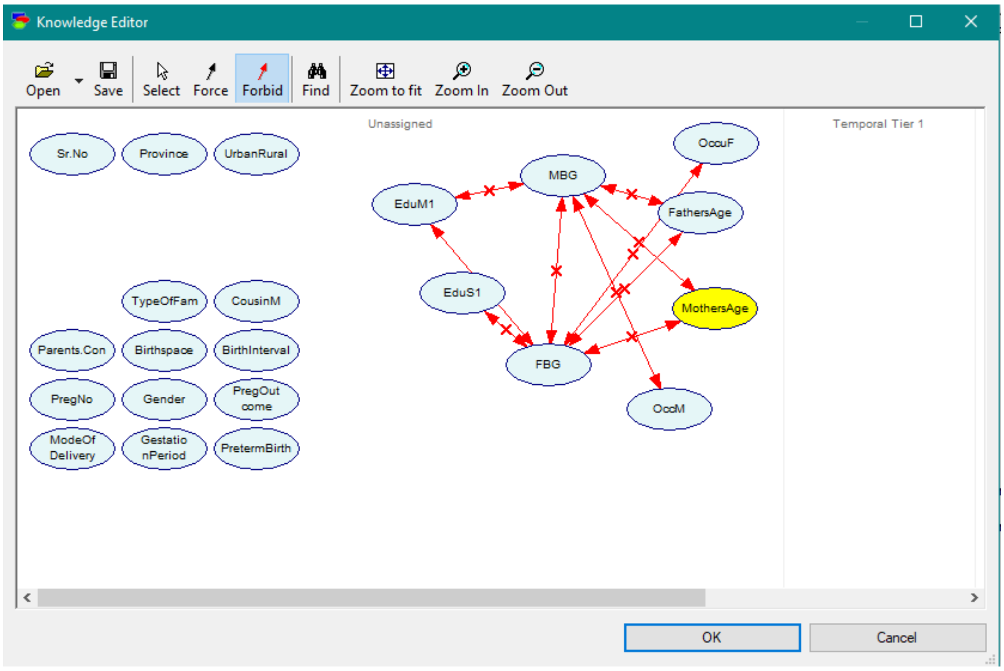


Figure 5.7: Background knowledge given by domain expert

5.2.3 Implementation of Association Rules

Association rules were applied in R by using package ’rattle’. Rattle is a graphical user interface and free open source library in R to implement different machine learning techniques [68]. It shows statistical summary of datasets. The dataset was uploaded in this library and parameters were set. The rules were then displayed on the interface. “Lift” was used as measurement criteria. If a rule has higher confidence but lower lift, that rule will seem more accurate because of its higher confidence. But accuracy of the rule independent of data can be misleading. The reason of using lift is that lift considers both the confidence of the rule and the overall dataset. More than 9000 rules were generated for the dataset used. Below is the flowchart of association rules.

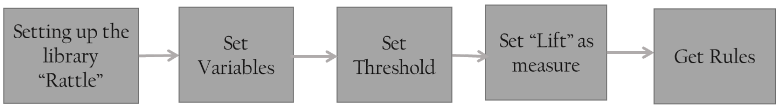


Figure 5.8: Flowchart of implementing Association Rules

5.3 Data Specification

The dataset has 30 factors in total. Feature selection was performed and 21 factors were selected. These factors are used as an input to our selected models. In R, factor is used to define feature in the dataset. Every factor has number of values associated with it, these values are called levels of that factor; such as the factor "Family type" has two factors ’Nuclear’ and ’Extended’. Nuclear means, a type of family where parents live with their kids while extended family is the one where grandparents, parents and kids live together. This means factor ’Family type’ consists of 2 factors. The factors of data is shown in Figure 5.9 along with their factor levels.

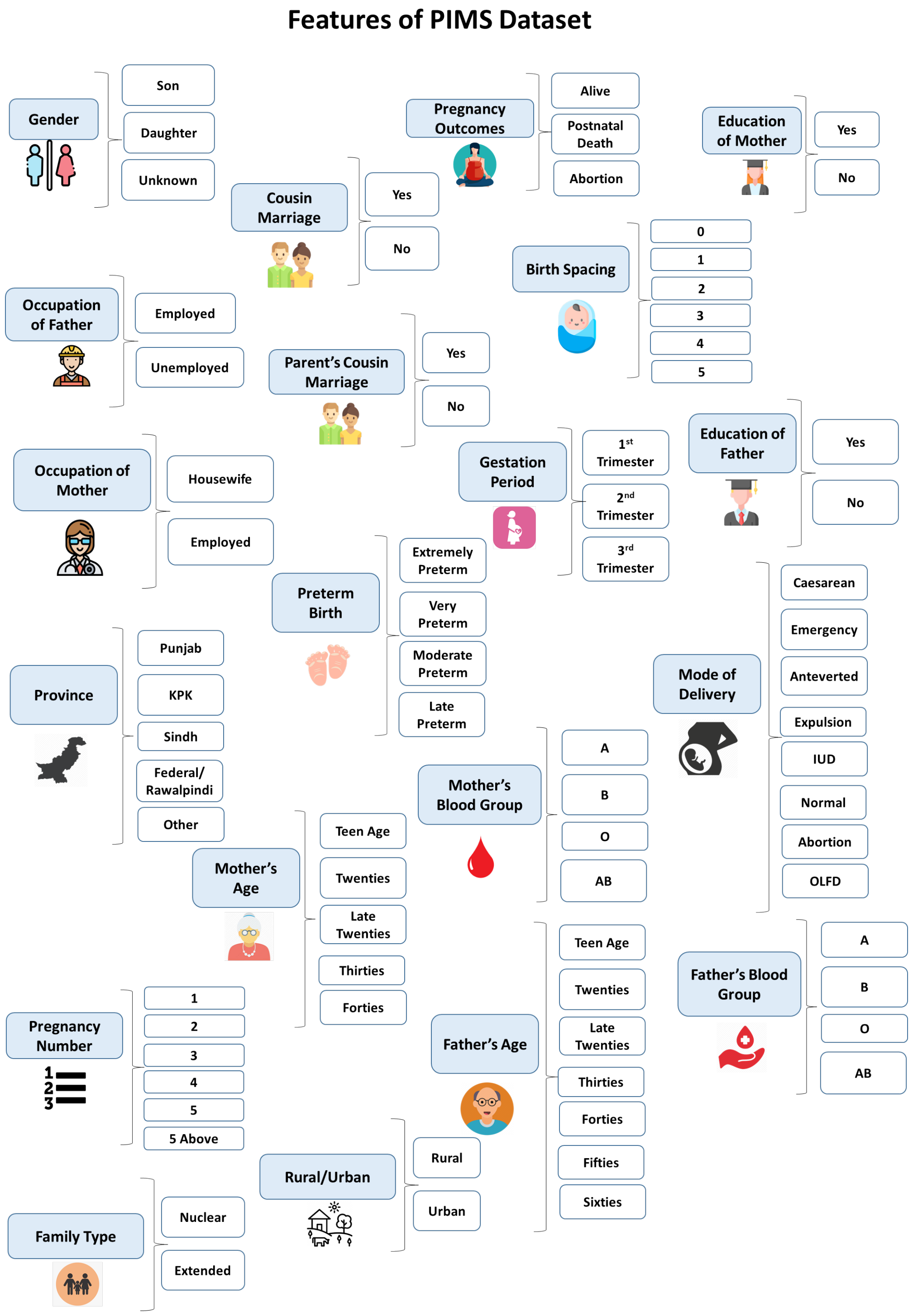


Figure 5.9: Factors of Data

5.4 Results and Discussion

5.4.1 Hidden Markov Model and Cousin Marriages

Hidden markov model has been extensively used in sequence analysis of observed events [69]. In this research, pregnancy events were analyzed.The pregnancy outcome such as abortion or miscarriage and gender of the child were observed keeping cousin marriage in account. How and if cousin marriage will impact the gender and miscarriages.

I. HMM graph for Cousin Marriage

Comparison of two models was performed: resulting model with cousin marriages and a model where subject was in non-cousin marriage is shown in figure 5.10 and figure 5.11 respectively. Graph shows:

- State 1 ,2 and 3 are the three observed states of this model
- The width of the edges shows the strength of that event to happen
- State 1 means 1st pregnancy event, State 2 indicates next pregnancy event and so on
- Straight color indicates the ratio of son, while diagonal line indicates the ratio of daughter and black color indicates the chances of miscarriages

The model shows that if State 1 occurs first the chances of having a daughter are more than the chances of having son. From State 1 there is 0.44 probability that State 2 will occur and 0.11 probability that State 3 will occur. Hence the chances of having State 2 occur after State 1 are more as the arrow indicates. State 2 shows that in the second pregnancy the chances of having a son and a daughter are almost equal with a slight chance of miscarriage. From State 2, i.e, second pregnancy the chance of State 3 to occur is only 0.11. If State 3 occurs in third Pregnancy then there is a high chance that a miscarriage will occur. From State 2, the probability of State 1 is higher to happen, i.e, 0.42 than State 3. Hence in terms of probability the flows of states will be as follows if State 1 happens first (Table 5.5):

Table 5.5: Flow of states with State 1 as the initial state (highest probability)

State1 (Initial State)	**State2**	**State1**	**State2**	**State1**
1 Pregnancy	2 Pregnancy	3 Pregnancy	4 Pregnancy	5 Pregnancy

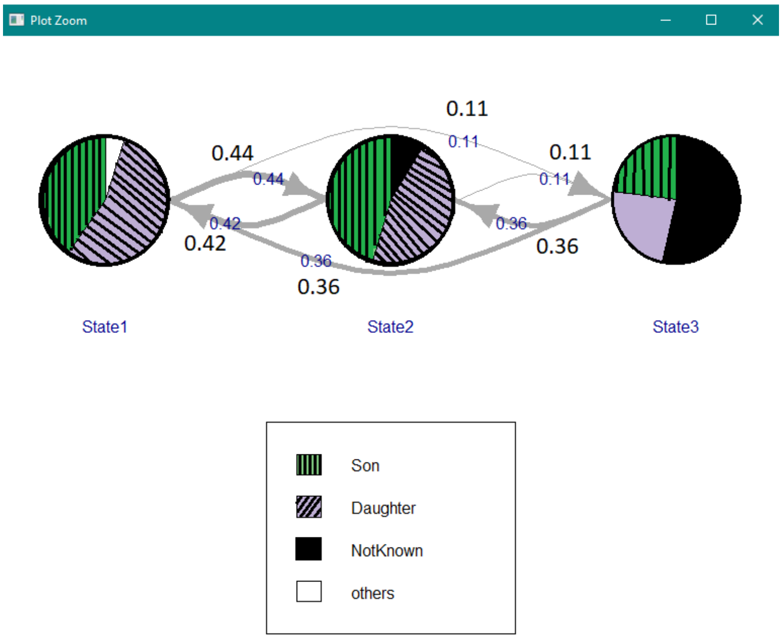


Figure 5.10: HMM with Cousin Marriage

If State 2 happens first, the flow of events can be seen in Table 5.6.

Table 5.6: Flow of states with State 2 as the initial state

State2 (Initial State)	**State1**	**State2**	**State1**	**State2**
1 Pregnancy	2 Pregnancy	3 Pregnancy	4 Pregnancy	5 Pregnancy

If State 3 happens first, the flow of events can be seen in Table 5.7.

Table 5.7: Flow of states with State 3 as the initial state

State3 (Initial State)	**State1/2**	**State1/2**	**State1/2**	**State1/2**
1 Pregnancy	2 Pregnancy	3 Pregnancy	4 Pregnancy	5 Pregnancy

II. HMM graph for Non-Cousin Marriage

Figure 5.11 shows the resulting model of non-cousin marriage. The model shows that in State 1 chance of having a daughter is comparatively higher than the chance of having a son. The flow of states will be as follows:

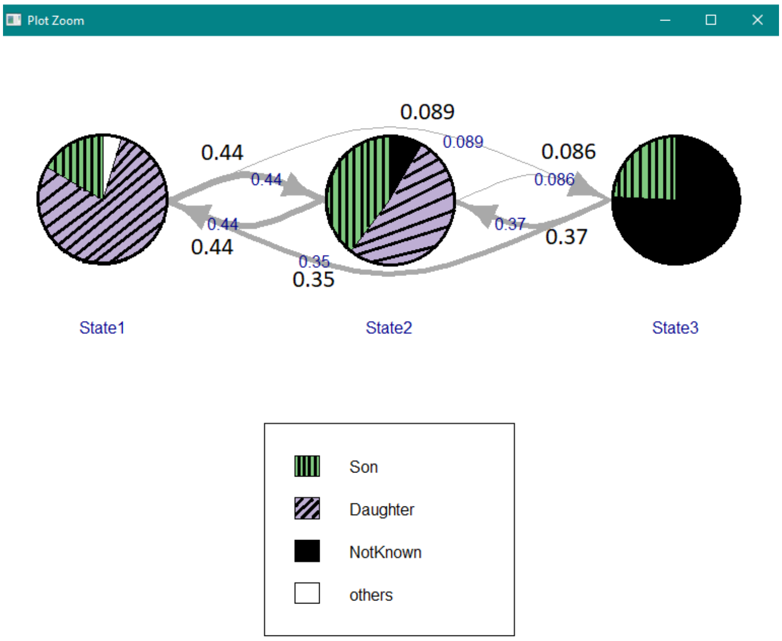


Figure 5.11: HMM with Non-Cousin Marriage

1. If State 1 happens first, chance of daughter is higher.
2. After State 1 there is 0.44 probability that State2 will occur. It means after 1st pregnancy, in the second pregnancy there is a slight chance of miscarriage and more chance of having a daughter than a son.
3. After State 2 there is a 0.44 probability that State 1 will occur and 0.086 probability that State 3 will occur.
4. If State 3 occurs there is a high chance that the pregnancy will end in miscarriage.
5. There is a 0.35 probability that the next pregnancy will follow the State1 i.e ratio of having a daughter is high than the ratio of having a son.

If State 1 happens first, the flow of events can be seen in Table 5.8

Table 5.8: Flow of states with State 1 as the initial state

State1 (Initial State)	**State2**	**State1**	**State2**	**State1**
1 Pregnancy	2 Pregnancy	3 Pregnancy	4 Pregnancy	5 Pregnancy

If State 2 happens first, the flow of events can be seen in Table 5.9

Table 5.9: Flow of states with State 2 as the initial state

State2 (Initial State)	**State1**	**State2**	**State1**	**State2**
1 Pregnancy	2 Pregnancy	3 Pregnancy	4 Pregnancy	5 Pregnancy

If State 3 happens first, the flow of events can be seen in Table 5.10. The flow of states is based on highest probability of that event to occur.

Table 5.10: Flow of states with State 3 as the initial state

State3 (Initial State)	**State2**	**State1**	**State2**	**State1**
1 Pregnancy	2 Pregnancy	3 Pregnancy	4 Pregnancy	5 Pregnancy

5.4.2 Comparison of Cousin and Non-Cousin Marriage Models

The models discussed above shows that in non-cousin marriages, there is a higher chance of miscarriages to happen as compared to cousin marriages. The rest of the events observed in both the models, chance of having a daughter in the first pregnancy is notably higher in non-cousin marriages as compared to cousin marriage. The rest of the flow of events were observed same in both the models.

5.4.3 Bayesian Network and analysis of pregnancy factors

By looking at the resulting learned structure as shown in figure 5.12, it can be seen that the factors that directly influence the pregnancy outcomes by applying bayes search are “Preterm Birth” and “Mode of delivery”. The indirect factors influencing the variable are “Birth Interval”, “Preg No” and “Gestation Period”. “Mother’s Age” is the tertiary factor that affects pregnancy outcomes. In Figure 5.12, the strength of influence shows how much a variable is affecting the other variable.

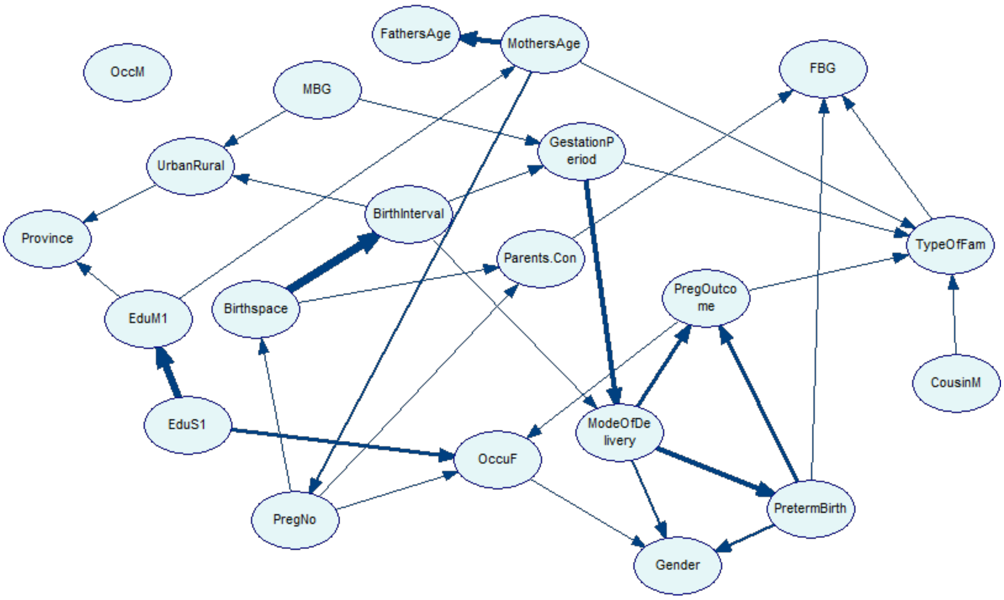


Figure 5.12: Strength of Influence

By looking at the learned structure, it can be seen that the factors that have high influence on pregnancy outcomes are:

- Mode of Delivery
- Preterm Birth
- Gestation Period
- Birth Interval

When setting the the influencing factors to the best parameters, alive births rates were observed to be 94%. this percentage was obtained by setting birth interval to "Optimal", preterm birth to "Late preterm", gestation period to "3rd trimester" and mode of delivery to "Normal" (Figure 5.13). In Figure 5.13, the bold blocks shows the variables with set parameters, pregnancy outcome is the variable under observation with 94% alive births.

The above results were for the Bayesian Search algorithm. The second algorithm applied was PC. Bayesian search follows the hill climbing procedure while PC performs classical independence tests. A comparative analysis of both the algorithms is then performed and best model is selected for the dataset on the basis of accuracy. By applying PC algorithm on the data set, it was found that the algorithm points out the same influential

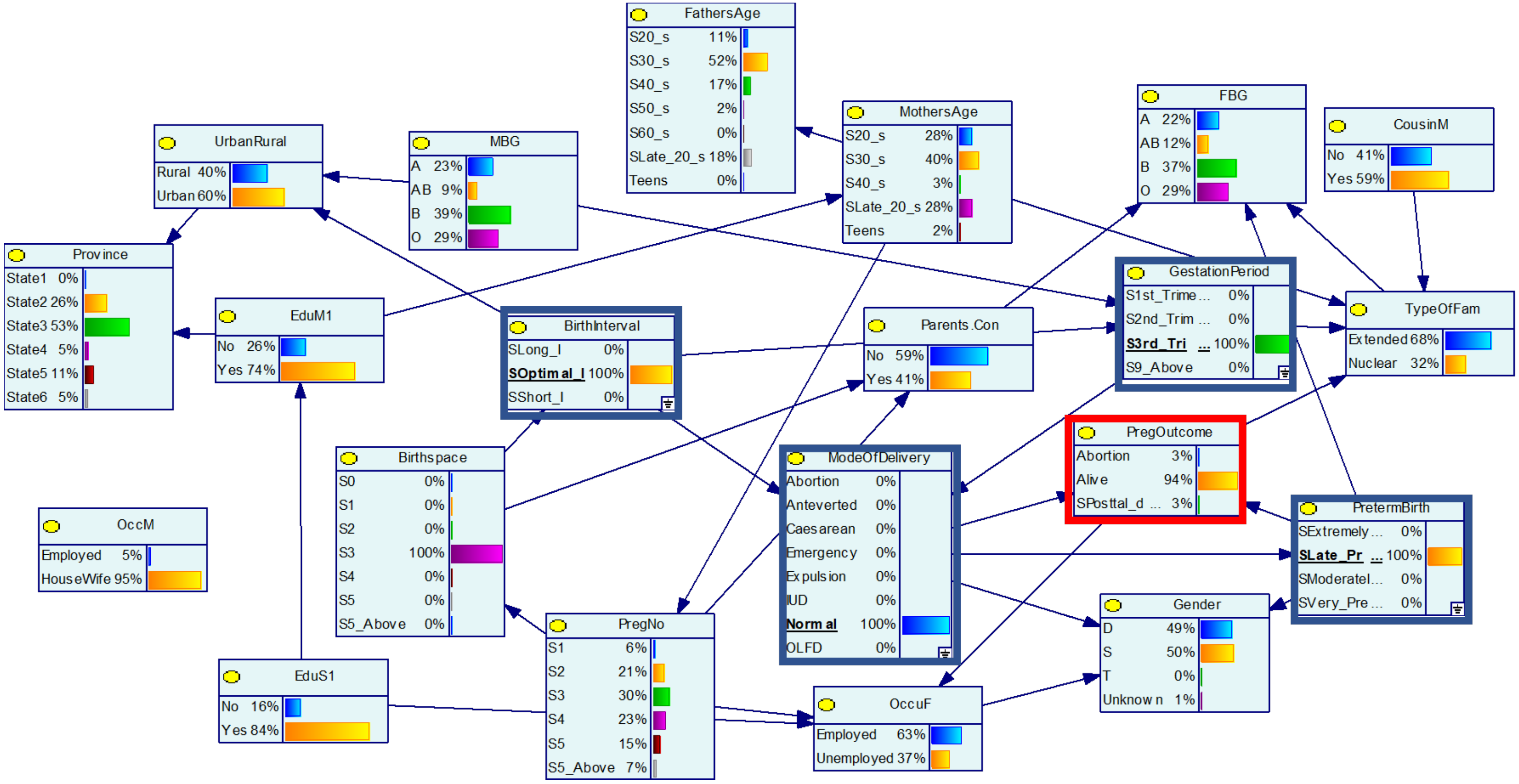


Figure 5.13: Setting the best parameters

factors for pregnancy outcome as produced by the resulting model of Bayesian search. Preterm birth, gestation period, birth interval, mother's age and pregnancy number were the factors that were identified by the PC algorithm too. Some new factors were also discovered, such as "Gender' was identified as a new direct influencing factor, mentioned below are some secondary influencing factors;

- Education of mother
- Place of residence (province)
- Subject's parent cousin marriage

Strength of influence shows (Figure 5.14 preterm birth, gender, birth interval to be most influencing factors on pregnancy outcome. The weight of arcs in given in the Table 5.11.

Table 5.11: Weight of different variables

Parent Node	**Child Node**	**Weight**
Gender	Pregnancy Outcome	0.343
Preterm Birth	Pregnancy Outcome	0.420
Gestation Period	Preterm Birth	0.310
Province	Preterm Birth	0.144
Birth Space	Preterm Birth	0.108
Preterm Birth	Mode of Delivery	0.108
Mother's Age	Birth Space	0.161
Education of Mother	Mode of Delivery	0.059
Subj. Parent's Consang.	Mode of Delivery	0.053

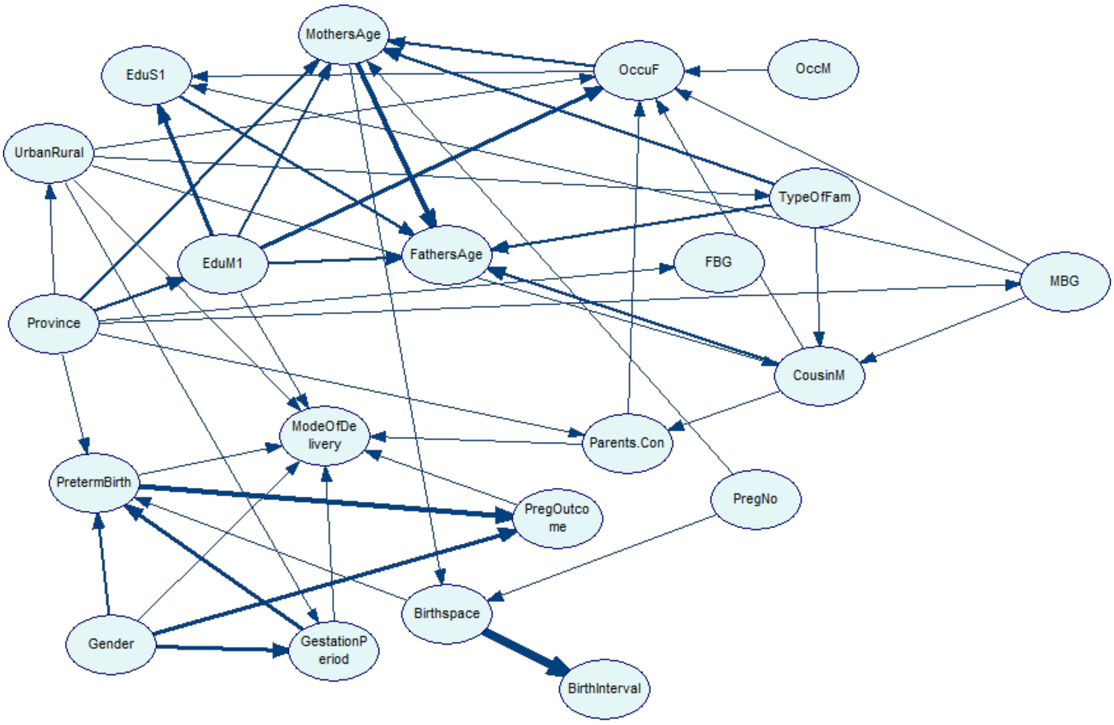
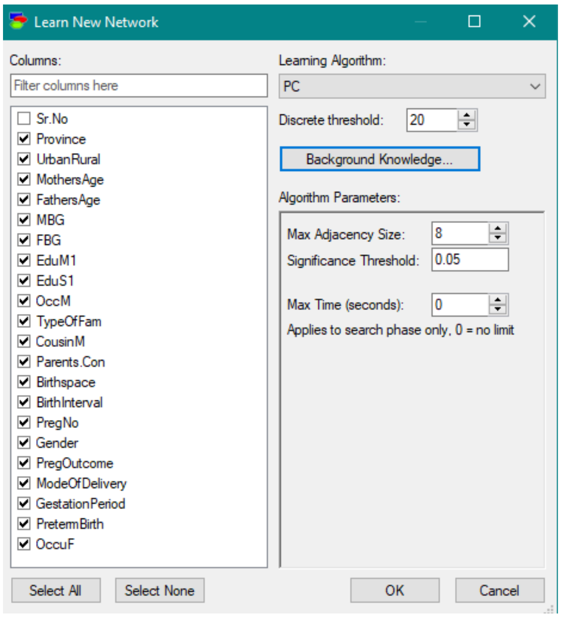


Figure 5.14: Strength of Influence for PC algorithm

By setting the influencing variables to the best parameters as shown in figure 5.16, 95% alive birth rates were achieved which was 84% before setting the parameters.



Columns:

Filter columns here

- □ Sr.No
- ☑ Province
- ☑ UrbanRural
- ☑ MothersAge
- ☑ FathersAge
- ☑ MBG
- ☑ FBG
- ☑ EduM1
- ☑ EduS1
- ☑ OccM
- ☑ TypeOfFam
- ☑ CousinM
- ☑ Parents.Con
- ☑ Birthspace
- ☑ BirthInterval
- ☑ PregNo
- ☑ Gender
- ☑ PregOutcome
- ☑ ModeOfDelivery
- ☑ GestationPeriod
- ☑ PretermBirth
- ☑ OccuF

Learning Algorithm: PC

Discrete threshold: 20

Algorithm Parameters:

Max Adjacency Size: 8

Significance Threshold: 0.05

Max Time (seconds): 0

Applies to search phase only, 0 = no limit

Figure 5.16: PC algorithm parameters

Hence an increase of 11% in alive birth rates were seen by setting the variables. this

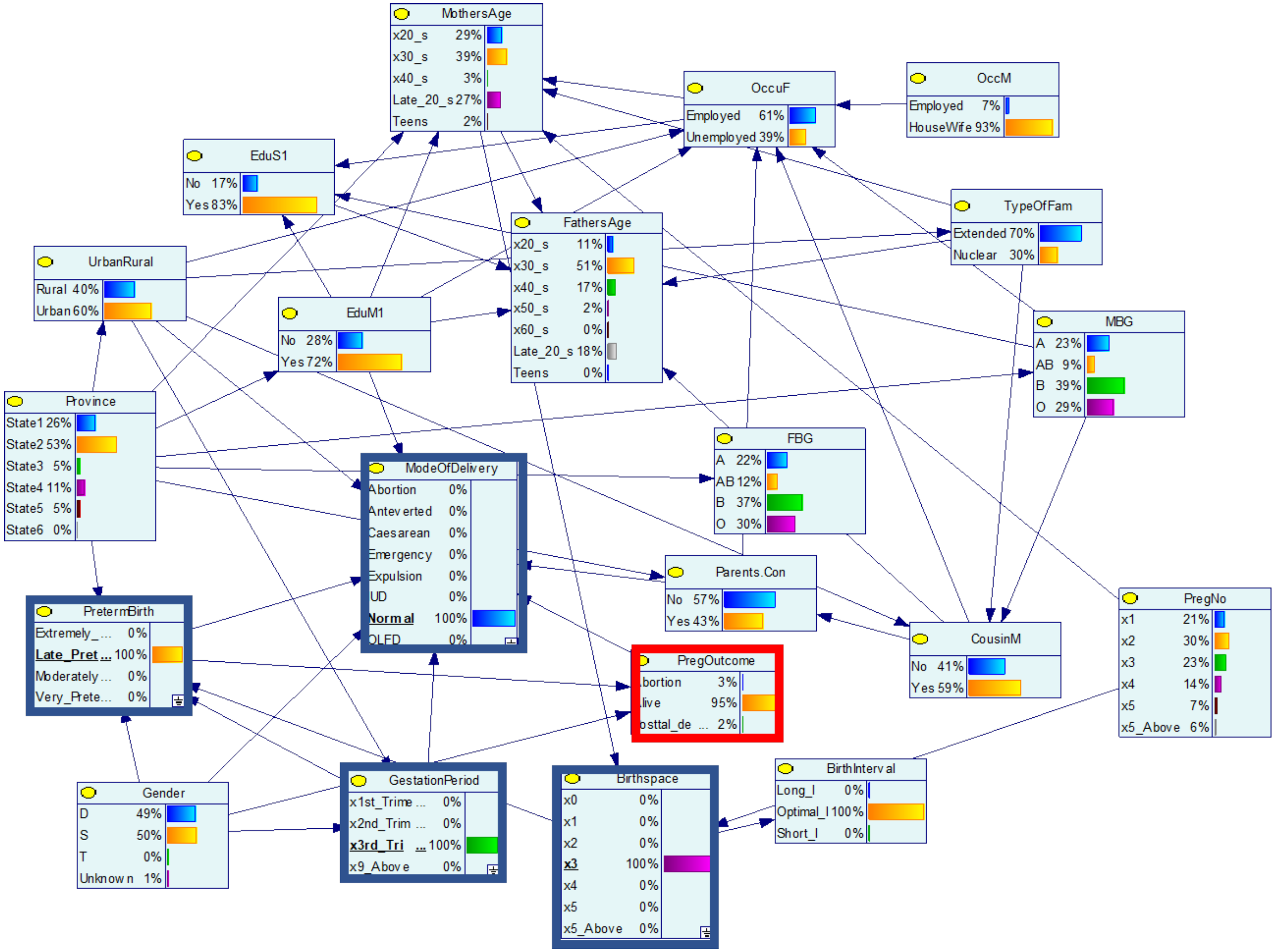


Figure 5.15: Setting the variable parameters for PC algorithm

percentage was obtained by setting birth interval to “Optimal”, preterm birth to “Late preterm”, gestation period to “3rd trimester” and mode of delivery to “Normal” (Figure 5.15). In Figure 5.15, the bold blocks shows the variables with set parameters, the pregnancy outcome is the variable under observation with 95% alive birth rate.

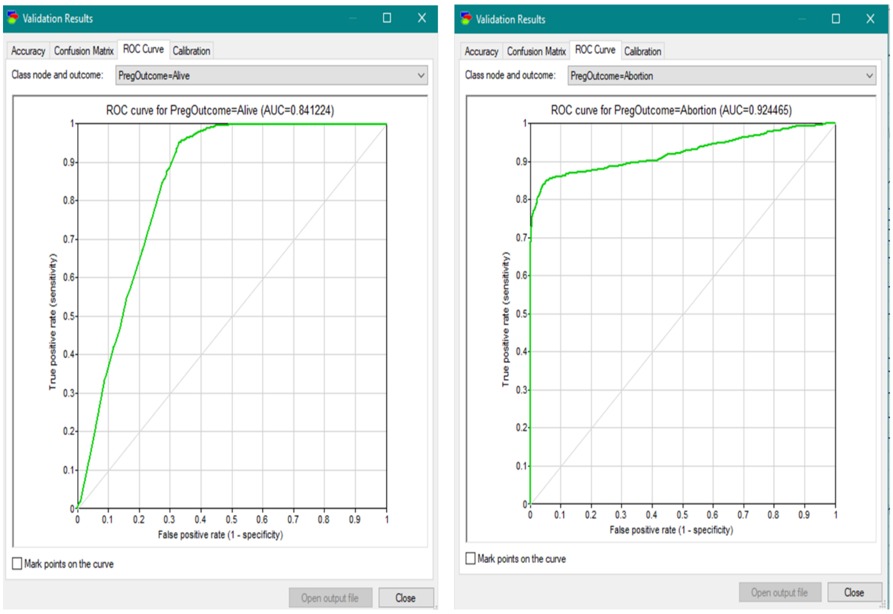
5.4.4 Comparison of PC and Bayesian Search

A comparison between the two algorithms was performed and on the basis of accuracy most authentic model will be considered as final model. The final conclusion will be based on the algorithm with higher accuracy. Below is the confusion matrix for Bayesian search algorithm:

Table 5.12: Confusion Matrix PC Algorithm

		Predicted		
		Abortion	Alive	Postnatal Death
Actual	Abortion	1282	424	7
Actual	Alive	59	10454	7
Actual	Postnatal Death	11	342	7

To analyze the performance of our model, AUC (Area under the curve) ROC (Receiver Operating Characteristics) was used. ROC analysis has become a popular method for evaluating the accuracy of medical diagnostic systems [70]. Its accuracy is not affected by decision criterion. AUC is widely used in measuring the effectiveness of diagnostic markers [71]. Hence model was evaluated using these techniques. Figure 5.17 and figure 5.18 shows the AUC curve for bayes search and PC algorithm respectively.



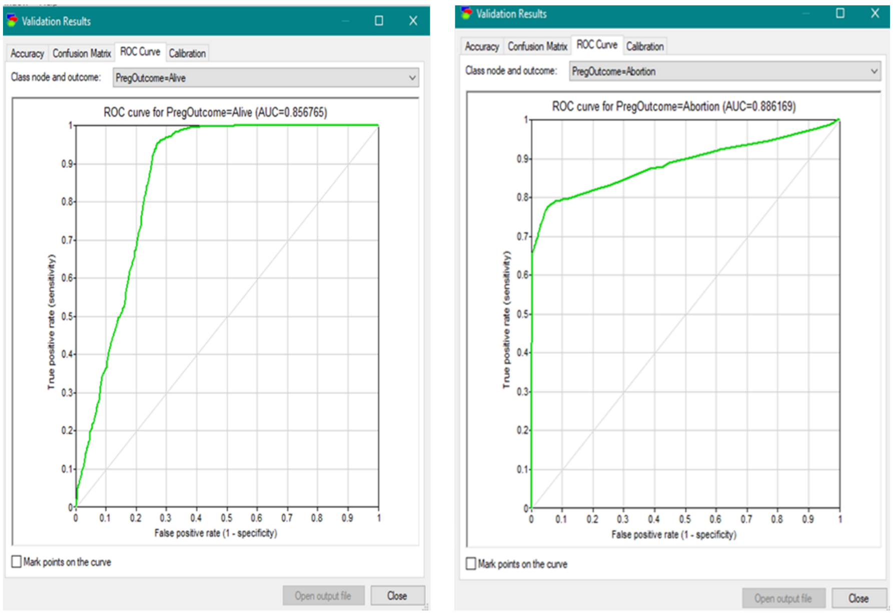
Class node and outcome: PregOutcome=Alive

□ Mark points on the curve

Class node and outcome: PregOutcome=Abortion

□ Mark points on the curve

Figure 5.17: AUC curve for Bayesian Search



Class node and outcome: PregOutcome=Alive

□ Mark points on the curve

Class node and outcome: PregOutcome=Abortion

□ Mark points on the curve

Figure 5.18: AUC curve for PC

5.4.5 Association Rules in pregnancy data set

An exploratory analysis was performed to analyze maternity dataset using data mining techniques. The objective of using this technique is to explore and identify multi-dimensional associations among different factors such as place of residence, birth interval, gestation period, pregnancy outcome, education of mother and father etc. Association rules were applied as studies shows that in the area of medical research association rule mining (ARM) can perform better than other techniques [72]. It is used while dealing with a different category of data and expresses hidden relationship between different variables [73]. As discussed in the previous chapter (see chapter 4) “Rattle” library in RStudio was used to implement association rules. “Pregnancy Outcome” was set as a target variable and “lift” was used as the criteria. Lift gives us the measure of importance of rule. More than 9000 rules were generated in total. Some of the interesting rules are shown in Table 5.13.

Table 5.13: List of Association Rules

Rule	Event	Lift score	Findings
R1	BirthInterval=Short-I,PregNo=1 ⇒ Birthspace = 1	1.67	Short birth intervals normally occur in 1st pregnancy
R2	BirthInterval=Short-I,Gestation Period = 3rd Trimester,Preterm Birth=Late Preterm ⇒ Birthspace=2	1.49	In birth interval of 2 years, birth normally completes its full term
R3	OccM=HouseWife, BirthInterval=Short-I, GestationPeriod=3rd Trimester, PretermBirth=Late Preterm ⇒ Birthspace=2	1.49	If wife stays at home, there is a chance of birth space of 2 years and birth will complete its full term
R4	EduM1=Yes, EduS1=Yes, TypeOfFam=Extended, BirthInterval=Short-I ⇒ Birthspace=1	1.43	When mother and father are educated and lives in an extended family, birth interval will be short i.e 1 year

R5	EduS1=Yes, TypeOfFam=Extended, Birth Interval=Short-I, Gestation Period = 3rd Trimester, PretermBirth = Late Preterm ⇒ EduM1=Yes	1.15	When father is educated and lives in an extended family, birth interval will normally be short and completes its full term, mother will likely to be educated
R6	EduM1=Yes, PregNo=1, PretermBirth=LatePreterm ⇒BirthInterval=Short-I	1.15	When mother is educated, and its 1st pregnancy, the birth will be late preterm and birth interval will be short
R7	EEduM1=Yes, TypeOfFam=Extended, GestationPeriod=3rd Trimester, Preterm Birth = LatePreterm, OccuF=Employed ⇒ EduS1=Yes	1.15	When mother is educated and father is employed, in an extended family, birth will mostly happen in 3rd trimester and father will likely to be educated
R8	FathersAge=30's, EduS1=Yes, OccM=HouseWife, Parents.Con=No ⇒ EduM1=Yes	1.15	When subject parents are not cousins and father is in 30's and educated, mother lives in house, the mother will likely to be educated
R9	UrbanRural=Urban, EduS1=Yes, OccM=HouseWife, TypeOfFam=Extended, BirthInterval=Short-I ⇒ EduM1=Yes	1.15	In urban areas, in extended family, where father is educated, mother is a house wife and birth intervals are short, mother will be educated

R10	EduM1=Yes,Preterm Birth = LatePreterm ⇒ ModeOfDelivery=Caesarean	1.15	In educated mothers with late preterm, the delivery will mostly be caesarean
R11	EduS1=Yes, Birthspace=1 ⇒ EduM1=Yes	1.15	When father is educated and birth space is 1, mother will be educated
R12	UrbanRural= Urban, EduM1 = Yes, OccM=HouseWife, Gestation Period=3rd Trimester ⇒ OccuF=Employed	1.15	In urban areas, where mother is educated and is a house wife, gestation period completes its 3rd trimester, fathers are mostly employed in such scenarios
R13	OccM=HouseWife, TypeOfFam=Extended ⇒ MothersAge=20's	1.14	In extended families where mother is a house wife, these mothers are mostly in their 20's
R14	EduS1=Yes, OccM=HouseWife, TypeOfFam=Extended, CousinM=Yes, BirthInterval=Short-I ⇒ EduM1=Yes	1.14	If father is educated, mother stays at home in an extended family and couple is close relative, short birth intervals are seen with educated mothers
R15	EduS1=Yes,Type Of Fam= Extended,Gender=D ⇒ EduM1=Yes	1.14	When father is educated in an extended family and gender of the baby is female, mother are seen to be educated

R16	EduM1=Yes,OccM = House Wife, TypeOfFam = Extended , CousinM=Yes, Birth Interval= Short-I ⇒ EduS1=Yes	1.14	When mother is educated and stays in home and lives in an extended family, where mother and father are cousins with short birth interval are seen, father happens to be educated
R17	OccM=HouseWife,CousinM=Yes, Birth Interval=Short-I, Gestation Period=3rd Trimester⇒TypeOfFam = Extended	1.07	In a cousin marriage, when mother stays at home with short birth intervals mostly birth occurs in 3rd trimester in an extended family
R18	EduM1=Yes,Parents.Con=Yes ⇒ CousinM=Yes	1.07	When girl is educated and parents are cousins, it is most likely that they marry their daughter in their family
R19	UrbanRural=Urban, OccM=HouseWife, GestationPeriod=3rd Trimester, OccuF=Employed ⇒ModeOfDelivery=Normal	1.07	In urban areas, where mother stays at home, birth occurs in 3rd trimester with normal delivery
R20	CousinM=Yes, Parents.Con=No ⇒TypeOfFam=Extended	1.07	In extended family, cousin marriage mostly takes place

R21	FathersAge=30's,
	EduS1=Yes,
	BirthInterval=Short-I⇒
	UrbanRural=Urban
		1.01	When father is educated and in his 30's he mostly prefer to have short birth intervals

R22	FBirthInterval=Short-I,
	ModeOfDelivery = Caesarean
	⇒ OccM=HouseWife
		1.01	In short birth interval, there is a high chance that delivery will be caesarean especially when mother stays at home

The list of rules mentioned in table 5.13 are only few to name, out of 9000 above rules given by the model. Some interesting facts were revealed using association rules, such as:

- R21: Fathers who are in their 30's prefer to have a short spacing between the births of their children.
- R22: Chances of caesarean delivery are higher for short birth intervals
- R19: When delivery happens in 3rd trimester, chances of normal delivery are high
- R1: In first pregnancy short birth intervals are observed
- R14: Short birth intervals are seen in extended family system even if mother is educated
- R18: Parents who are cousins tend to marry their children in their families
- R17: In cousin marriages, where short birth intervals and 3rd trimester births are observed, this pattern is mostly seen in extended families

Not all rules given by the algorithm hold value and give interesting information. Out of 9000 rules extracted from the dataset, most valued rules were picked out , which has to be done manually and hence its tiresome [18]. The table 5.13 shows that we cannot deduct valuable information from all the rules. Hence we need to drop those rules that don't show interesting patterns.

The interesting thing about this research is that all the techniques that were applied on the given data set are totally different from each other, in terms of their uses and implementation. But the information gathered from all the models backs each other's findings. The patterns found by Hidden Markov Model were validated by Bayesian network. The influential variables found using Bayesian model was validated by the two algorithms applied on the data set. Moreover AUC ROC curve showed the authenticity of the Bayesian model. Frequent patterns were found using association rules and interesting rules were discovered. Hence research objectives were achieved with >90% authenticity.

Chapter 6

Conclusion and Future Work

This chapter will discuss briefly about the research that was conducted and contributions of the research objectives on social and academic level. This research work found the influential factors such as birth interval, gestation period, cousin marriage and some others effects on the pregnancy events and frequent patterns in PIMS dataset. Conclusion and contribution were discussed in section 6.1 and 6.2 respectively. In section 6.3, limitations of the work and future work related to the problem statement are discussed.

6.1 Conclusion

Neonatal mortality is a major global issue nowadays. Pakistan is among those five countries that are accounted for half of all newborn's death. Many researches have been carried out around the world to find out the major causes of these deaths but many of them suffers from lack of data dimensionality, biasness in data and small size of data. These approaches also failed to analyze causality among factors. Wwithout studying causality the root cause of the issue can not be identified. In this work the maternity factors were analyzed such as birth interval, gestation period, pregnancy number, cousin marriages and many others while examining the causality among them and used the data set with almost no biasness and high level of dimensionality. With the help of proposed models PIMS data was explored in all those perspectives. This reserach was able to fill the gap found in existing studies (see section 3.6). The proposed model developed to analyze different factors gave an accuracy of 92%.

6.2 Contribution

With the advancement in technology people wonder if it will contribute in healthcare as much as it is contributing towards other fields. This research was initiated with the believe that introduction and widespread use of data mining and machine learning in healthcare will play an important role in life-saving. The opportunities are virtually limitless for the technology to improve and accelerate healthcare center, workflow, and financial outcomes. The research will contribute towards mothers and neonates healthcare in following ways:

1. **Machine Learning and Data Analytics in Maternal and Child Health care:** This research indicates that Bayesian network models, hidden markov models and association rules can also be used to solve and predict medical health problems related to mothers and neonates. As the models gave a better accuracy score i.e above 90% hence this shows the authenticity of our models and gives a confidence to use these models in making of health care systems.

2. **Prevent Neonatal Deaths and improved mothers health:** This research may help in reducing mortality rates by predicting the factors associated with neonatal deaths. Doctors can monitor these factors and intervene to reduce that risk by focusing on patient-specific risk factors. Knowing the influential factors, the doctors will educate the mothers, this will help in reduction of mortality rates. Preventing those factors that affects neonates health and taking precautionary measures during and after pregnancy will help in lowering the rates of Pakistan. Data analytics and Machine learning can improve health of expecting mothers in a targeted, efficient, and patient-centered manner. Once the doctor and health care workers know the real cause of high neonatal mortality rates, they can provide guidance to the mothers. The mothers once gets aware of the risk factors associated with pregnancy will maintain their health accordingly.

3. **Achieving Millennium Development Goal 4 (MDG 4):** The target of MDG4 is to reduce under five mortality rate by two-thirds 2/3 during the period of 1995-2015 [74], which Pakistan failed to achieve. Pakistan is found to be 83% off track towards MDG4 [75]. In Southern Asia (India, Pakistan, Bangladesh) many deaths occurred due to preventable causes which could be avoided with simple,

high-impact, and cost-effective interventions [76].

6.3 Limitations

The model has given better results as compared to those approaches used in previous researches. The size of data set can be improved. More data from different regions of the country can be added to the proposed model to make the system more efficient. As more data is available, better information can be provided to the system.

6.4 Future Work

Data from different hospitals around the country can be added in the proposed model in future, to make the system robust. Moreover, we applied bayesian network, hidden markov model and association rules for analyzing PIMS dataset. In future algorithms such as naive bayes, support vector machines, random forest, principal component analysis (PCA) and other machine learning techniques can be applied to improve the results.

References

[1] Amanullah Khan, Mary V Kinney, Tabish Hazir, Assad Hafeez, Stephen N Wall, Nabeela Ali, Joy E Lawn, Asma Badar, Ali Asghar Khan, Qudsia Uzma, et al. Newborn survival in pakistan: a decade of change and future implications. *Health policy and planning*, 27(suppl_3):iii72–iii87, 2012.

[2] Sidney B Westley and Minja Kim Choe. How does son preference affect populations in asia? 2007.

[3] World Health Organization. *World health statistics 2016: monitoring health for the SDGs sustainable development goals.* World Health Organization, 2016.

[4] Poh Lin Tan, S Philip Morgan, and Emilio Zagheni. A case for "reverse one-child" policies in japan and south korea? examining the link between education costs and lowest-low fertility. *Population research and policy review*, 35(3):327–350, 2016.

[5] Therese Hesketh and Jiang Min Min. The effects of artificial gender imbalance. *EMBO reports*, 13(6):487–492, 2012.

[6] Sadia Saeed. Toward an explanation of son preference in pakistan. *Social Development Issues*, 37(2):17–36, 2015.

[7] Khaula Atif, Muhammad Zia Ullah, Afeera Afsheen, Syed Abid Hassan Naqvi, Zulqarnain Ashraf Raja, and Saleem Asif Niazi. Son preference in pakistan; a myth or reality. *Pakistan journal of medical sciences*, 32(4):994, 2016.

[8] Claire Gudex, Bentt Løwe Nielsen, and Monika Madsen. Why women want prenatal ultrasound in normal pregnancy. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 27(2):145–150, 2006.

[9] Fareeha Ubaid, Erum Shahani, Farah Saleh, et al. Reasons for disclosure of gender to pregnant women during prenatal ultrasonography. *International journal of women's health*, 5:781, 2013.

[10] Asifa Kamal and Muhammad Khalid Pervaiz. Factors affecting the family size in pakistan: Clog-log regression model analysis. *Journal of Statistics*, 18(1), 2011.

[11] Muhammad Zaman. Marriage of cousins: Congenital diseases and people's perceptions in pakistan, a public health challenge. *Journal of public health policy*, 31(3): 381–383, 2010.

[12] Sarah Bundey and Hasina Alam. A five-year prospective study of the health of children in different ethnic groups, with particular reference to the effect of inbreeding. *European Journal of Human Genetics*, 1(3):206, 1993.

[13] Aatekah Owais, Abu Syed Golam Faruque, Sumon K Das, Shahnawaz Ahmed, Shahed Rahman, and Aryeh D Stein. Maternal and antenatal risk factors for stillbirths and neonatal mortality in rural bangladesh: a case-control study. *PloS one*, 8(11):e80164, 2013.

[14] Samir B Kassar, Ana MC Melo, Sônia B Coutinho, Marilia C Lima, and Pedro IC Lira. Determinants of neonatal death with emphasis on health care during pregnancy, childbirth and reproductive history. *Jornal de pediatria*, 89(3):269–277, 2013.

[15] S Farrokh Mostafavi. Estimating the causal effect of maternal education on infant mortality with dhs data for iran. In *Marrakech International Population Conference*, volume 2, 2009.

[16] Yavar Naddaf, Mojdeh Jalali Heravi, and Amit Satsangi. Predicting preterm birth based on maternal and fetal data. *Google Scholar*, 2008.

[17] Sarah Rabbani and Abdul Qayyum. Comparative analysis of factor affecting child mortality in pakistan. *Research Journal Social Sciences*, 4(2):1–17, 2017.

[18] Kangmoon Kim and Young-Mee Lee. Understanding uncertainty in medicine: concepts and implications in medical education. *Korean journal of medical education*, 30(3):181, 2018.

[19] Finn V Jensen et al. *An introduction to Bayesian networks*, volume 210. UCL press London, 1996.

[20] Beáta Reiz and Lehel Csató. Bayesian network classifier for medical data analysis. *International Journal of Computers Communications & Control*, 4(1):65–72, 2009.

[21] Rivera.R. Strengths and weaknesses of hidden markov models, 1996.

[22] SS Ravi David C. Torney Srinivas Doddi, Achla Marathe. Discovery of association rules in medical data. *Medical informatics and the Internet in medicine*, 26(1): 25–33, 2001.

[23] Jyoti Soni, Ujma Ansari, Dipesh Sharma, and Sunita Soni. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8):43–48, 2011.

[24] Mervi Eerola and Satu Helske. Statistical analysis of life history calendar data. *Statistical Methods in Medical Research*, 25(2):571–597, 2016.

[25] Satu Helske, Fiona Steele, Katja Kokko, Eija Räikkönen, and Mervi Eerola. Partnership formation and dissolution over the life course: applying sequence analysis and event history analysis in the study of recurrent events. *Longitudinal and Life Course Studies*, 6(1):1–25, 2014.

[26] Satu Helske and Jouni Helske. Mixture hidden markov models for sequence data: the seqhmm package in r. *arXiv preprint arXiv:1704.00543*, 2017.

[27] Satu Helske, Jouni Helske, and Mervi Eerola. Analysing complex life sequence data with hidden markov modelling. In *LaCOSA II: Proceedings of the International Conference on Sequence Analysis and Related Methods*. LIVES-Swiss National Centre of Competence in Research; Swiss National . . . , 2016.

[28] Michal Horny. Bayesian networks. *Boston University, Boston*, 2014.

[29] SS Ravi David C. Torney Srinivas Doddi, Achla Marathe. Discovery of association rules in medical data. *Medical informatics and the Internet in medicine*, 26(1): 25–33, 2001.

[30] Stephen E Brossette, Alan P Sprague, J Michael Hardin, Ken B Waites, Warren T Jones, and Stephen A Moser. Association rules and data mining in hospital infection

control and public health surveillance. *Journal of the American medical informatics association*, 5(4):373–381, 1998.

[31] Antje Horsch, Leah Gilbert, Stefano Lanzi, Ji Kang, Yvan Vial, and Jardena Puder. Associations between maternal stress during pregnancy and obstetric and neonatal outcomes. *Psychoneuroendocrinology*, 83:28, 09 2017. doi: 10.1016/j.psyneuen.2017.07.313.

[32] BN Lakshmi, TS Indumathi, and Nandini Ravi. A study on c. 5 decision tree classification algorithm for risk predictions during pregnancy. *Procedia Technology*, 24:1542–1549, 2016.

[33] RHF Van Oppenraaij, E Jauniaux, OB Christiansen, JA Horcajadas, RG Farquharson, and N Exalto. Predicting adverse obstetric outcome after early pregnancy events and complications: a review. *Human reproduction update*, 15(4):409–421, 2009.

[34] Marjorie R Sable and Deborah Schild Wilkinson. Impact of perceived stress, major life events and pregnancy attitudes on low birth weight. *Family planning perspectives*, pages 288–294, 2000.

[35] Stephen E Brossette, Alan P Sprague, J Michael Hardin, Ken B Waites, Warren T Jones, and Stephen A Moser. Association rules and data mining in hospital infection control and public health surveillance. *Journal of the American medical informatics association*, 5(4):373–381, 1998.

[36] Gulam Muhammed Al Kibria, Vanessa Burrowes, Allysha Choudhury, Atia Sharmeen, Swagata Ghosh, Arif Mahmud, and KC Angela. Determinants of early neonatal mortality in afghanistan: an analysis of the demographic and health survey 2015. *Globalization and health*, 14(1):47, 2018.

[37] Yu Mu, Kai Feng, Ying Yang, and Jingyuan Wang. Applying deep learning for adverse pregnancy outcome detection with pre-pregnancy health data. In *MATEC Web of Conferences*, volume 189, page 10014. EDP Sciences, 2018.

[38] Christiana R Titaley, Michael J Dibley, Kingsley Agho, Christine L Roberts, and John Hall. Determinants of neonatal mortality in indonesia. *BMC public health*, 8 (1):232, 2008.

[39] Perianayagam Arokiasamy and Abhishek Gautam. Neonatal mortality in the empowered action group states of india: trends and determinants. *Journal of biosocial science*, 40(2):183–201, 2008.

[40] Jocelyn E Finlay, Melanie K Norton, and Iván Mejía Guevara. Adolescent fertility and child health: The interaction of maternal age, parity and birth intervals in determining child health outcomes. *International Journal of Child Health and Nutrition*, 6(1):16–33, 2017.

[41] Yasir Bin Nisar and Michael J Dibley. Determinants of neonatal mortality in pakistan: secondary analysis of pakistan demographic and health survey 2006–07. *BMC Public Health*, 14(1):663, 2014.

[42] W Henry Mosley and Lincoln C Chen. An analytical framework for the study of child survival in developing countries. *Population and development review*, 10(0): 25–45, 1984.

[43] Fareeha Ubaid, Erum Shahani, Farah Saleh, et al. Reasons for disclosure of gender to pregnant women during prenatal ultrasonography. *International journal of women's health*, 5:781, 2013.

[44] Khaula Atif, Muhammad Zia Ullah, Afeera Afsheen, Syed Abid Hassan Naqvi, Zulqarnain Ashraf Raja, and Saleem Asif Niazi. Son preference in pakistan; a myth or reality. *Pakistan journal of medical sciences*, 32(4):994, 2016.

[45] Sonia Omer, Sitwat Farooq, and Sadia Jabeen. Effects of cousin marriages on adverse pregnancy outcomes among women in pakistan: A secondary analysis of data from the pakistan demographic and health survey 2012-13. *Pakistan Journal of Women's Studies= Alam-e-Niswan= Alam-i Nisvan*, 23(1):65, 2016.

[46] A Mushfiq Mobarak, Theresa Chaudhry, Julia Brown, Tetyana Zelenska, M Nizam Khan, Shamyla Chaudry, Rana Abdul Wajid, Alan H Bittles, and Steven Li. Estimating the health and socioeconomic effects of cousin marriage in south asia. *Journal of biosocial science*, 51(3):418–435, 2019.

[47] Fawaz Amin Saad and Eric Jauniaux. Recurrent early pregnancy loss and consanguinity. *Reproductive biomedicine online*, 5(2):167–170, 2002.

[48] Sonia Omer, Sitwat Farooq, and Sadia Jabeen. Effects of cousin marriages on adverse pregnancy outcomes among women in pakistan: A secondary analysis of data from the pakistan demographic and health survey 2012-13. *Pakistan Journal of Women's Studies= Alam-e-Niswan= Alam-i Nisvan*, 23(1):65, 2016.

[49] Shrikant Kuntla, Srinivas Goli, TV Sekher, and Riddhi Doshi. Consanguineous marriages and their effects on pregnancy outcomes in india. *International Journal of Sociology and Social Policy*, 33(7/8):437–452, 2013.

[50] Jasim Anwar, Siranda Torvaldsen, Mohamud Sheikh, and Richard Taylor. Under-estimation of maternal and perinatal mortality revealed by an enhanced surveillance system: enumerating all births and deaths in pakistan. *BMC public health*, 18(1): 428, 2018.

[51] Rahim Moineddin, Flora I Matheson, and Richard H Glazier. A simulation study of sample size for multilevel logistic regression models. *BMC medical research methodology*, 7(1):34, 2007.

[52] Magdalena Babińska, Jerzy Chudek, Elżbieta Chełmecka, Małgorzata Janik, Katarzyna Klimek, and Aleksander Owczarek. Limitations of cox proportional hazards analysis in mortality prediction of patients with acute coronary syndrome. *Studies in Logic, Grammar and Rhetoric*, 43(1):33–48, 2015.

[53] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

[54] James Monroe. Sql joins explained. *Sql-join.*

[55] Sean R Eddy. What is a hidden markov model? *Nature biotechnology*, 22(10):1315, 2004.

[56] Dimitris Margaritis. Learning bayesian network model structure from data. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2003.

[57] WJ Frawley et al. Knowledge discovery in databases: An overview", knowledge discovery in databases, piatetsky-shapiro and frawley (eds.), aaai, 1991.

[58] Smith D. R is hot?, 2010.

[59] Alexis Gabadinho, Gilbert Ritschard, Matthias Studer, and Nicolas S Müller. Mining sequence data in r with the traminer package: A user's guide. 2009.

[60] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

[61] Zachary SL Foster, Thomas J Sharpton, and Niklaus J Grünwald. Metacoder: An r package for visualization and manipulation of community taxonomic diversity data. *PLoS computational biology*, 13(2):e1005404, 2017.

[62] S Lauritzen. Genie modeler, 2010.

[63] KA Wilson, DD Wallace, SS Goudar, D Theriaque, and EM McClure. Identifying causes of neonatal mortality from observational data: A bayesian network approach. In *Proceedings of the International Conference on Data Mining (DMIN)*, page 132. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2015.

[64] Mark Voortman, Denver Dash, and Marek J Druzdzel. Learning causal models that make correct manipulation predictions with time series data. In *Causality: Objectives and Assessment*, pages 257–266, 2010.

[65] David Heckerman. A bayesian approach to learning causal networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 285–295. Morgan Kaufmann Publishers Inc., 1995.

[66] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach.* Malaysia; Pearson Education Limited,, 2016.

[67] Kumar Ravi and Sheopujan Singh. Bayesian network for uncertainty representation in semantic web: a survey. *International Journal of Computer Applications Technology and Research*, 2(5):530–538, 2013.

[68] Graham Williams. *Data mining with Rattle and R: The art of excavating data for knowledge discovery.* Springer Science & Business Media, 2011.

[69] Richard Hughey and Anders Krogh. Hidden markov models for sequence analysis: extension and analysis of the basic method. *Bioinformatics*, 12(2):95–107, 1996.

[70] Karimollah Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2):627, 2013.

[71] David Faraggi and Benjamin Reiser. Estimation of the area under the roc curve. *Statistics in medicine*, 21(20):3093–3106, 2002.

[72] Rohini R Rao and Krishnamoorthi Makkithaya. Identifying risk patterns in public health data through association rules. *The Journal of BMESI*, pages 30–34, 2016.

[73] Cheikh Ndour, Aliou Diop, and Simplice Dossou-Gbété. Classification approach based on association rules mining for unbalanced data. *arXiv preprint arXiv:1202.5514*, 2012.

[74] Farah Asad Mansuri. Situation analysis of millennium development goals 4 & 5: Pakistan's perspective. *Annals of Abbasi Shaheed Hospital & Karachi Medical & Dental College*, 19(2), 2014.

[75] Farah Asad Mansuri. Situation analysis of millennium development goals 4 & 5: Pakistan's perspective. *Annals of Abbasi Shaheed Hospital & Karachi Medical & Dental College*, 19(2), 2014.

[76] Anil B Deolalikar. Attaining the millennium development goals in pakistan: How likely and what will it take to reduce infant mortality, child malnutrition, gender disparities and to increase school enrollment and completion?, 2005.