# Predictive Modeling Using Decision Tree for Low Birth-Weight

Md Rezwan Islam, Gaurav Sitaula, Ali Shakiba

Group 13

# Executive Summary

The primary objective of this report is to build a predictive model based on the decision tree model using the low birth weight dataset which contains data from more than seventeen thousand observations. Low birth weight is one of the fatal reasons behind the child mortality rate all over the world. Therefore, a successful decision tree which could predict the low birth weight of an infant would enable to handle this issue in an effective way.

To build the model, 37 variables are chosen as the predictor variables which entails the socio-economics, behavioral aspect of parents, the health status of the mother and other related issues. The predictor variable is the low birth weight of a child which is a binary variable as well. As the decision tree takes care of missing values using its internal algorithm, no special treatment is applied here. Thereafter, the dataset is partitioned into two sets: a) training b) validation using 50:50 split rule. Subsequently, serval decision tree models are built to compare with each other to choose the best model. Based on the misclassification rate, the best model contains nine leaves with the misclassification error of 35.45%.

In conclusion, this type decision tree would predict the low birth rate of the infants which in real sense could bring a positive change for the nations all over the world who are suffering this low birth weight problem.

## Introduction

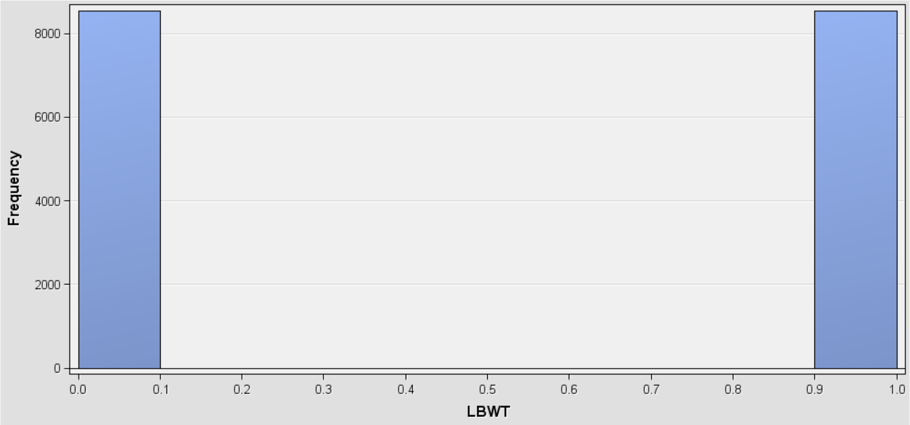
Low birth weight is considered as one of the fatal causes behind child mortality all over the world. According to the WHO's report, "More than 20 million infants worldwide, representing 15.5 per cent of all births, are born with low birthweight, 95.6 per cent of them in developing countries" (Wardlaw, Blanc, Zupan, & Åhman, 2004). In addition to that, the report mentions that low birthweight is closely associated with fetal and neonatal mortality and morbidity, inhibited growth and cognitive development, and chronic diseases later in life. (Wardlaw, Blanc, Zupan, & Åhman, 2004). Therefore, the reduction of low birth weight might contribute to the decrease of child mortality as well as other complicated related diseases.

Keeping this context in mind, the purpose of this report is to develop a predictive modelling using Decision Tree to understand the vital factors and thereby predict the probability of an infants with low birth weight. Successful model would enable the healthcare providers and the government to tackle this issue and to plan.

The first step of building a data model is to understand the dataset and prepare the dataset for data modelling. In the following segment, we would discuss the data set and then the preparation of data.

## Exploratory Data Analysis

This dataset contains 17,097 records of which 50% have low birth weight and 50% have normal birth weight. Birthweight is affected by several factors which includes parents socio-economic and behavioral conditions, prior pregnancy related data, mother's own diet, body condition and so on and so forth. To examine these factors, 37 predictor variables are chosen and data are collected on these variables. Out of these 37 predictor variables, 13 are interval, 21 are nominal and rest 3 variables are binary variables. The Response Variable of this data set is Low birth weight baby, defined as weight less than 2500 grams. This is defined as 1= Low birth weight vs 0 = normal birth weight which indicates it as a binary response variable. The following graph shows the histogram of the LBWT.

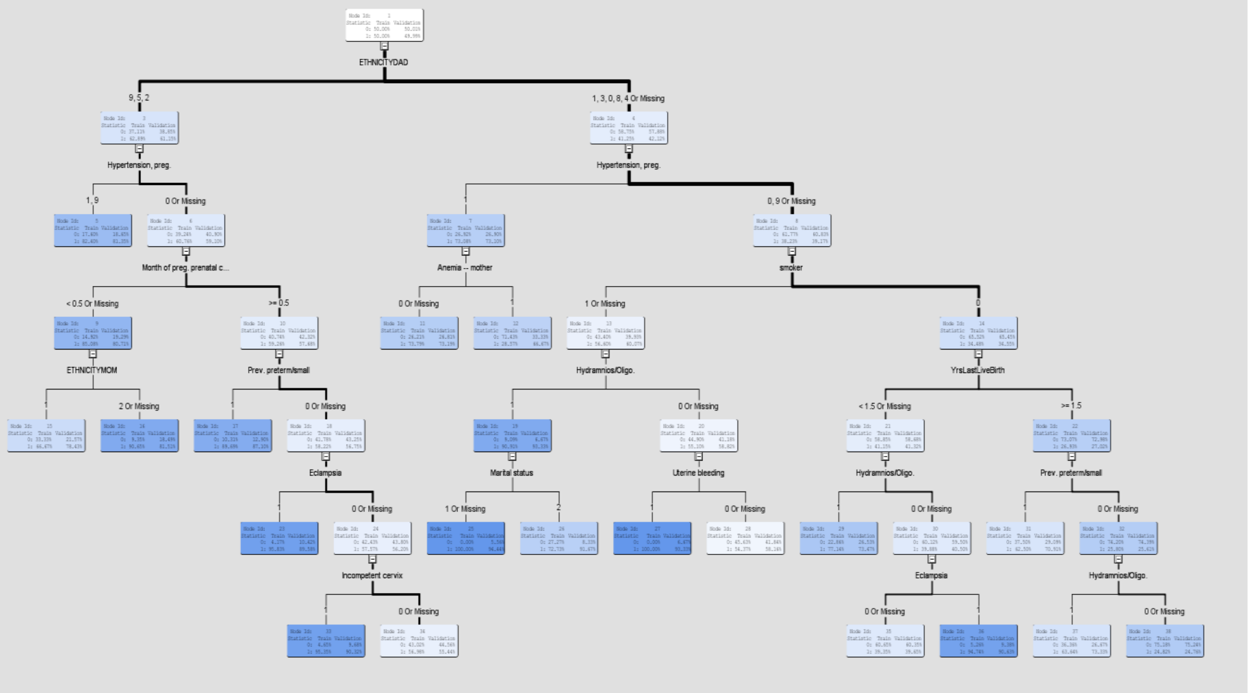


## Data Preparation

In the second step of analysis, we first consider the missing values and their treatments. As the decision tree algorithm inherently deals with the missing values, no special treatment is applied to the those values. Thereafter, we have divided our dataset into two separate data sets. One is the training dataset and another is the validation data set. We have followed 50:50 split criterion which means fifty percent of the data goes to the training data set and rest 50 percent of the data goes to the validation data set. Technically, we have created a data partition node, connected it to the data source and then run the program in the SAS Enterprise Miner. By doing so, the data set is ready to build any model upon it and test the result of the prediction on the validation data that is created by the split.

## Model Building

In the third step, we have first build the maximal decision tree model. Maximal model refers the maximum number of splits or leaves that the algorithm could generate. This tree contains 19 leaves with average square error of 0.21962 and misclassification rate of 0.354778 of the validation data set. Here is the tree diagram of the maximal model.



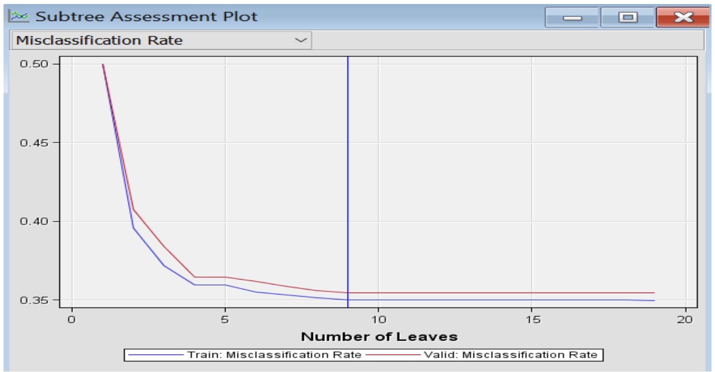
Based on this maximal tree, our next step of analysis is to obtain the best tree. In other words, our objective is to gain a decision tree which has minimal number of leaves and optimum amount of misclassification rate and average square errors. To achieve our desired tree, we have built five separate decision tree based on 1) all four-available subtree assessment measurement 2) by increasing the maximum branch from two to three. Thereafter, we have compared all our five model to obtain the best one. In general, the basic idea behind this process is to prune the maximal tree based on separate criteria so that both dataset would show us the optimal amount of leaves and desired level of subtree assessment measurement. Below is the comparison chart of our five intermediate data models that we have built.

| Name of the decision tree | Validation Misclassification Rate | Validation Average Square Error |
|---|---:|---:|
| maximal tree | 0.354778337 | 0.219609526 |
| Optimal Tree_ASE | 0.354544391 | 0.219277658 |
| Optimal tree_Lift | 0.354544391 | 0.219277658 |
| Optimal Tree_3 branch_Dec | 0.355714119 | 0.225566637 |
| optimal tree_Decision | 0.354544391 | 0.225136369 |
| Optimal tree_Missclass | 0.354544391 | 0.225136369 |

Based on the validation misclassification rate, we have four models which have the similar misclassification rate which is the smallest at the same time. On the other hand, based on the validation average square error, we have two models that have the lowest amount of error. Following table shows number of leaves each of these trees have.

| Name of the decision tree | Number of Leaves |
|---|---:|
| maximal tree | 19 |
| Optimal Tree_ASE | 16 |
| Optimal tree_Lift | 16 |
| Optimal Tree_3 branch_Dec | 12 |
| optimal tree_Decision | 9 |
| Optimal tree_Missclass | 9 |

We assume that primary concern in our predictive model is to minimize misclassification rate. Here, we should trade-off between our average squared errors with the number of leaves to simply the model within the tolerable range. As the number of leaves increase, the complexity of the decision tree increases but the error decreases. From the above table, it is evident that the number of leaves decreases from 19 to 9 but the average squared error has not increased to a significant amount whereas the misclassification rate remains. Therefore, we have decided to pick our best model with 9 leaves and having misclassification rate of 0.354 which is shown below:



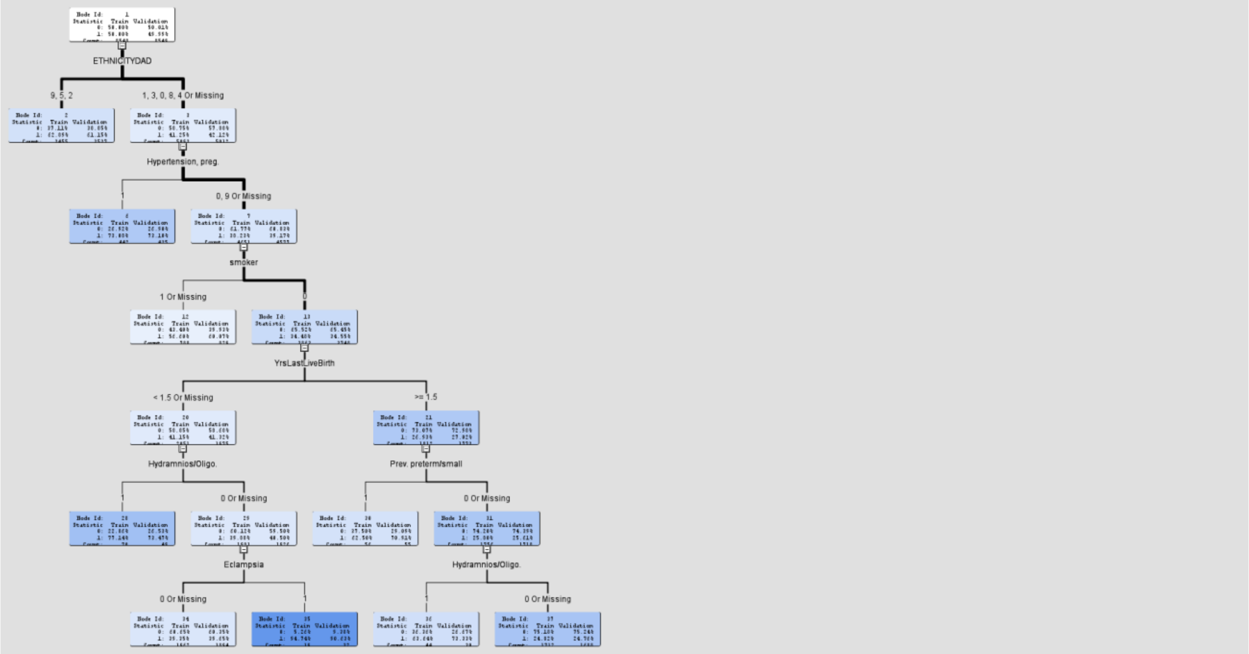
Important variables for this best decision tree is given below.

```
Variable Importance

                                                                                 Ratio of
                                         Number of                              Validation
                                         Splitting                Validation    to Training
Variable Name       Label                  Rules      Importance  Importance    Importance

ETHNICITYDAD                                 1          1.0000      1.0000        1.0000
HYPERPR             Hypertension, preg.      1          0.7130      0.7875        1.1045
smoker                                       1          0.5763      0.7694        1.3352
YrsLastLiveBirth                             1          0.4492      0.5091        1.1335
HYDRAM              Hydramnios/Oligo.        2          0.4055      0.3991        0.9843
PRETERM             Prev. preterm/small      1          0.2754      0.3782        1.3736
ECLAMP              Eclampsia                1          0.2447      0.3332        1.3615
```

Here is the best decision tree:



Eventually, if we have the inputs of the important predictor variables, then based on this tree we would be able to predict the low birth weight of an infant with 35.45% misclassification error.

## Bibliography

- Wardlaw, T., Blanc, A., Zupan, J., & Åhman, E. (2004). *Low Birth Weight: Country, Regional and Global Estimates* (1st ed.). New York, NY 10017: United Nations Children’s Fund and World Health Organization.

- Wardlaw, T., Blanc, A., Zupan, J., & Åhman, E. (2004). *Low Birth Weight: Country, Regional and Global Estimates* (1st ed.). New York, NY 10017: United Nations Children’s Fund and World Health Organization.