Investigating Determinants of Birth Weight Using Bayesian Tree-Based Nonparametric Modeling

by

Adam Michael Kurth

A Thesis
Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved May 2025 by the
Graduate Supervisory Committee

P. Richard Hahn, Chair
Shuang Zhou
Shiwei Lan

ARIZONA STATE UNIVERSITY
August 2025

ABSTRACT

Low birth weight (LBW) remains a critical public-health indicator, linked strongly with higher neonatal mortality, developmental delays, and lifelong chronic diseases. Using the 2021 U.S. Natality dataset (> 3 million births), this thesis develops a Bayesian, tree-based, nonparametric framework that models the full birth-weight distribution and quantifies LBW risk.

The raw dataset is condensed into 128 mutually exclusive classes defined by seven dichotomous maternal-infant predictors and 10 (or 11) birth-weight categories, comprised of 10% LBW quantile categories and one additional aggregated normal birth-weight (NBW) category. The full and LBW-only models are grown to contrast and investigate how variable selection is altered based on the restriction the dataset. The models are Classification and Regression Trees (CART) using the marginal Dirichlet-Multinomial likelihood as the splitting criterion. This criterion is equipped to handle sparse observations, with the Dirichlet hyperparameters informed by previous quantiles from the 2020 dataset to avoid "double dipping."

Employing a two-tier parametric bootstrap resampling technique, a 10,000 tree ensemble is grown yielding highly stable prediction estimates. Maternal race, smoking status, and marital status consistently drive the initial LBW risk stratification, identifying black, smoking, unmarried mothers among the highest-risk subgroups. When the analysis is restricted to LBW births only, infant gender and maternal age supersede smoking and marital status as key discriminators, revealing finer biological gradients of risk. Stable and informative mean ensemble estimates are obtained with narrow 95% percentile intervals.

The resulting modeling framework combines the interpretability of decision trees with a custom quasi-Bayesian splitting criterion, yielding delivering actionable, clinically relevant insights for targeting maternal-health interventions among the most vulnerable subpopulations.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, mentor, and committee chair, Dr. P. Richard Hahn, whose guidance, support, and expertise have been instrumental throughout my research journey. His depth of knowledge in Bayesian statistics and tree-based models has shaped the foundation of this work. As both a mentor and a friend, his example is one I aspire to emulate in my own career.

I am sincerely thankful to my committee members, Dr. Shuang Zhou and Dr. Shiwei Lan, for their flexibility, insightful feedback, and thoughtful contributions, all of which have enriched the depth and rigor of this thesis. In particular, I would like to thank Dr. Zhou for her inspiring teaching and intellectual curiosity, which played a pivotal role in both my undergraduate and master's success.

Special thanks to the School of Mathematical and Statistical Sciences at Arizona State University for fostering an environment of academic excellence and growth. I am especially grateful for the sense of community and support that the School provides—an environment that has shaped not only my academic path but also my personal development.

I am deeply grateful to my father for instilling in me a lifelong intellectual curiosity and passion for learning, and for being the earliest source of encouragement in pursuing mathematics and statistics. My family unwavering support, especially during periods of serious illness has been foundational to every step of my academic journey. I am especially indebted to my mother for her steadfast love, consistency, and strength during those most difficult times. Her support made my success possible.

Lastly, I acknowledge the researchers whose foundational work on low birth weight determinants and Bayesian nonparametric methods laid the groundwork for this study. This research stands on their shoulders.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Background & Problem

Low birth weight (LBW), defined as a birth weight less than 2.5 kilograms (Cutland, C. L., Lackritz, E. M., Mallett-Moore, T., Bardají, A., Chandrasekaran, R., Lahariya, C., Nisar, M. I., Tapia, M. D., Pathirana, J., Kochhar, S., Muñoz, F. M., Brighton Collaboration Low Birth Weight Working Group, 2017; Kramer, 1987), is a significant public health indicator. Infants born with LBW face substantially higher risk for neonatal mortality, developmental delays, and chronic health problems such as respiratory and neurological impairments (Finch, 2003). These adverse outcomes arise from a complex interplay of genetic, biological, environmental, and socioeconomic factors. Understanding and identifying determinants of LBW is therefore critical for developing informing targeted preventative measures and improving neonatal outcomes.

Decades of research confirm that LBW is a multifactorial issue. An early landmark meta-analysis by Kramer (1987) reviewed 895 studies from 1974-1984 and identified 43 causal determinants. Kramer concluded that maternal anthropometry (height and pre-pregnancy weight), inadequate gestational weight gain, cigarette smoking, malaria infection, and a history of adverse pregnancy outcomes exert *independent* effects on intrauterine growth restriction (IUGR), while few factors influence gestational duration (Kramer, 1987). The highly-interrelated nature of these risk factors led to confounding, yielding an impediment for modeling birth-weight outcomes by their interactive, not additive, effects. For example, inadequate pregnancy weight gain might depend on whether she smokes or has health conditions. Classical regression models such as logistic regression, typically assume

additive effects and thus miss such interactions. As a result, traditional models often struggle to disentangle which combinations of maternal-infant characteristics *truly* signify an at-risk pregnancy.

Subsequent work in epidemiology repeatedly show that these risk factors are not found in isolation. In 2006, Kitsantas *et al.* (2006) use Classification and Regression Trees (CART) developed by Breiman *et al.* (1984) to identify high-risk profiles of LBW on a large dataset of Florida birth records, uncovering important context-specific combinations that influence LBW risk. For instance, mothers who smoked *and* had inadequate weight-gain during pregnancy had sharply elevated LBW risk. This study highlights the strengths of interpretability using CART, but still lacks predictive power over logistic regression by relying entirely on empirical observations. Moreover, reducing the problem to a binary LBW indicator variable discards vital information about how far a newborn falls below the LBW threshold. Regardless of differences, a baby just above 2.5 kg is treated the same as a much heavier baby, and all LBW cases are treated alike. This binary cutoff masks important differences in the birth-weight distribution.

Recent research has moved beyond classification toward estimating the full birth-weight distribution conditional on covariates. Bayesian nonparametric mixture models allow the birth-weight density to vary flexibly across subpopulations defined by maternal factors, without strong parametric assumptions (Dunson *et al.*, 2008). Other approaches use copula-based or density regression techniques to jointly model birth weight with related outcomes such as gestational age (Rathjens *et al.*, 2024). These methods can capture detailed distributional effects of predictors, such as how covariates influence the entire left tail of the birth-weight distribution. However, a drawback in these advanced models is their complexity and lack of interpretability for users. In contrast, practitioners particularly in public health often prefer models that yield clear, simple decision rules for identifying high-risk subgroups.

Several demographic and socioeconomic factors are well-known to influence LBW risk. Younger and older mothers are associated with higher incidence of LBW (Goisis *et al.*, 2017). Lower educational attainment is associated with limited health care access (Finch, 2003; Jain, 2024), and environmental exposures, including tobacco use, substance abuse, and air pollution, further elevate risk by interfering with fetal growth and development (Stanford Medicine, 2025; Lu *et al.*, 2020). Inadequate prenatal care is another important factor of LBW outcomes (Institute of Medicine, 1985). Crucially, LBW incidence also varies sharply by racial and economic contexts. In the United States, the LBW incidence rate for Black infants is double that of white newborns, comparing 14.7% to 7.1% (March of Dimes, 2024). These patterns underscore the multifactorial nature of effects that influence LBW outcomes and the need to account for diverse influences in any predictive model.

Taken together, these considerations highlight a central challenge in birth-weight modeling: existing methods trade flexibility for interpretability. Approaches using decision trees alone provide transparent subgroup rules while ignoring the full birth-weight distributions, simultaneously advanced Bayesian density models capture distributional details but lack intuitive clarity. This work aims to bridge the gap by developing a quasi-Bayesian tree-based framework that stratifies the population into interpretable risk subpopulations while modeling full birth-weight distributions for predicting LBW outcomes.

Chapter 2

LITERATURE REVIEW & METHODOLOGY

Previous Work on Birth-Weight Modeling

As previously mentioned, Kramer (1987)'s meta-analysis identified 43 LBW determinants then categorized them into genetic, nutritional, psychosocial, etc. and assessed their effects on birth weight and prematurity. Maternal profiles were separated based on income status, for high-income mothers, smoking status, poor maternal nutrition or low pre-pregnancy weight were the strongest LBW determinants whereas in low-income settings, maternal race origin, undernutrition, short stature, and malaria exposure were found to be the most important predictors (Kramer, 1987). While for preterm births, smoking status and low pre-pregnancy weight are strong indicators (Kramer, 1987). Kramer (1987) concluded by stating that many potential contributors remain under studied, naming maternal work, prenatal care, and previous infections as some examples. This comprehensive seminal work highlights the complex and multifactorial nature of LBW, leaving open questions about interactions of factors and distributional outcomes, motivating a more flexible modeling procedure.

The application of CART by Kitsantas *et al.* (2006) to 181,690 singleton births from Florida, led the identification of high-risk LBW mothers. Known risk factors of smoking status, gestation weight-gain, parity, etc. were used to grow separate decision trees by geographic region and compared against logistic regression (Kitsantas *et al.*, 2006). The CART model revealed high-risk profiles for White and Hispanic mothers with low pregnancy weight gain, parity, and marital status defined high-risk stratification among non-smokers (Kitsantas *et al.*, 2006). For instance, smoking mothers that gain less than

20 lbs are at significantly higher risk than mothers of larger weight-gain pregnancies, and Black mothers form a high-risk subpopulation in some regions regardless of other factors (Kitsantas *et al.*, 2006). However, predictive accuracy was marginally better than logistic regression (Kitsantas *et al.*, 2006), the recursive partitioning procedure conducted by CART uncovered some of the complex factor interaction in the LBW data. This study shows how the order of factors could be useful in disentangling strong interaction effects, suggesting room for improved or alternative methods.

Dunson *et al.* (2008) (2008) used Bayesian semiparametric methods to link maternal pregnancy weight gain to birth-weight distributions. Using a Dirichlet-process mixture, they flexibly defined clusters of women by their weight-gain trajectories and jointly modeled birth-weight densities across clusters (Dunson *et al.*, 2008). This approach allowed the *entire* birth-weight distribution to vary with weight-gain patterns, including distribution tails, while also capturing heterogeneity of how pregnancy factors influence birth-weight (Dunson *et al.*, 2008). Dunson et. al. demonstrated that modeling the full distribution in perinatal data is insightful — beyond mean estimates. However, advanced Bayesian models, latent clustering, and complex MCMC procedures lack interpretability and are computationally intensive, highlighting the need for model simplicity while retaining flexibility.

More recently, Rathjens *et al.* (2024) in 2024 proposed a Bayesian distribution regression approach using copulas to jointly model birth weight and gestational age. Marginal distributions are assumed to follow a Gaussian for birth-weight outcomes, Dagum distribution for skewed gestational age, and the copula linked the two cumulative distribution functions (CDF) as functions of covariates (Rathjens *et al.*, 2024). The results of this study show non-linear effects of gestational age on weight and tail-dependent associations were captured by a Clayton copula (Rathjens *et al.*, 2024). The focus of bivariate outcomes here show how distribution modeling can extend traditional regression approaches. Beyond complex copula models, Bayesian methods enrich perinatal risk modeling.

In 2024, Jain (2024) proposed a scalable Bayesian density estimation method for nationally collected birth records. Inspired by kernel density methods, a Gaussian mixture is employed to model conditional distributions of birth weights given various predictors. Through advanced MCMC and targeted subsampling techniques, the model was able to capture complex patterns and estimate birth-weight densities at scale. Jain's work estimates the full distribution by density regression but underscores the computational and interpretational challenges.

Current Approach & Contributions

In this thesis, we adopt CART and Bayesian nonparametric methods to approximate birth-weight distributions. CART is a nonparametric algorithm proposed by Breiman *et al.* (1984) and implemented in R by Therneau et al. (Therneau and Atkinson, 2023), called *Recursive Partitioning and Regression Trees*, or `rpart`. The algorithm works in two stages: tree construction and tree pruning.

First, the tree is constructed. Given the data, `rpart` recursively partitions it into binary splits on the given predictor variables, creating nodes at each split. Though the splits need not be binary, this provides a clear and interpretable tree. CART employs a greedy approach to building decision trees (Centre for Speech Technology Research, nd), where its goal is to maximize homogeneity or equivalently minimize heterogeneity in the data. At each node, CART evaluates all possible splits on candidate predictors and chooses the one that best "explains" the data by minimizing the node impurity, resulting in two child nodes with more homogeneous subgroups (Therneau and Atkinson, 2023). This process is applied recursively to each child node then growing a larger tree until the tree's max depth is reached or no further improvement is found (Therneau and Atkinson, 2023).

Once fully grown, the tree typically overfits to the data, yielding large errors for small fluctuations. To address this, cross-validation is used to estimate prediction error for a

sequence of pruned trees (Therneau and Atkinson, 2023). The tree is then "trimmed" back to the best cross-validation performance (Therneau and Atkinson, 2023), yielding the final tree that balances complexity and accuracy. For each terminal node (or "leaf") in the final tree, a sequence of if-then conditions categorize birth-weight outcomes based on maternal covariates.

Interpretability is preserved by using CART to automatically uncover high- and low-risk groups for subpopulations of specific maternal and infant characteristics, much akin to Kitsantas *et al.* (2006). Additionally, in line with Dunson *et al.* (2008) and Jain (2024), this tree-based method imposes no strict distributional assumptions on the birth-weight responses allowing for nonlinear interactions and heterogeneous effects to be captured naturally by CART. Our preprocessing procedure results in count data of various birth-weight categories, motivating the use of the *marginal Dirichlet-Multinomial (DM) likelihood* as the Bayesian "evidence" and splitting criterion. The DM likelihood is chosen by producing posterior predictive distributions and interval estimation at each leaf whereas the Gini index measures only impurity. The impurity of a terminal node, is entirely dependent on the sample size by relying on empirical proportions (stats.stackexchange.com, 2025b). Additionally, the Gini index is known to suffer with data sparsity (Kamperis, 2021). Compared to normal birth-weight (NBW) observations, we expect a large discrepancy between the total number of observed LBW and NBW counts.

Birth-weight count observations can be safely assumed to follow a *multinomial* distribution, and the Dirichlet prior smooths categories not observed. For use in CART, a split with high-DM likelihood translates as added improvement from parent to child nodes, reducing heterogeneity. The DM likelihood will be derived formally in Section 2 as a favorable splitting criterion for our application.

Overview of the Birth-Weight Dataset

The primary dataset for this analysis is the 2021 Vitality Statistics Natality Birth Data (National Bureau of Economic Research, 2024). Collected by the National Center for Health Statistics (NCHS), this dataset contains a detailed record of birth outcomes and various maternal characteristics as part of the Vital Statistics Cooperative Program (Jain, 2024; National Bureau of Economic Research, 2024). Standing as one of the most comprehensive datasets with over 3 million birth-weight records for maternal and infant health in the United States, collected annually across all states and District of Columbia since 1972 (National Bureau of Economic Research, 2024).

For this analysis, variables in the 2021 data are broadly categorized into three domains: demographic, health, and geographic. Demographic features include date of birth, parental age and education, marital status, birth order, sex, and geographic location. Health features cover birth weight, gestational age, prenatal care adequacy, delivery attendants, and Apgar scores, while geographic indicators include state, county, and metropolitan status (National Bureau of Economic Research, 2024). Note that Apgar scores are examinations based on newborn vitals five minutes following delivery, observing how newborns handle being outside the mother's womb (apg, nd).

Data Preprocessing & Feature Engineering

The preprocessing procedure transforms the high-dimensional 2021 dataset into a workable and condensed dataset for computational efficiency, while preserving key information about predictors. Preprocessing involved (1) encoding all categorical and continuous variables into unique dichotomous predictors, (2) dimension reduction from 3 million rows to 128 unique predictor combinations, and (3) creating a consolidated counts dataset, primarily expanding the LBW region by creating birth-weight categories based on quantile cut

points. From the dataset, seven key predictor variables and birth-weight outcomes (in kg) are retained for modeling.

Binary Feature Encoding

To enhance interpretability and computational efficiency, only seven predictors are selected based on clinical relevance, strong generalizability, and prior research support. The encoding procedure was inherited from Jain (2024), and these predictors serve as an example of a small, yet representative set of predictors. Note that the encoding of `mrace15` is suggested by Jain (2024) and March of Dimes (2024) as the primary dichotomy, *though this choice is entirely arbitrary*. According to 2024 U.S. Census Bureau (2024), the national population is roughly 75.3% White and 13.7% Black, which provides demographic context for this binary split. All information of each feature representation and meaning is conveyed in the table below.

Table 2.1: Binary predictor definitions used in this study

Label	**Natality field**	**Value = 1**	**Value = 0**
Boy	`sex`	Infant is male (“M”)	Infant is female (“F”)
Married	`dmar`	Mother is married	Mother not married
Black	`mrace15`	Black / African American	Any other race
Over33	`mager`	Maternal age $>$ 33 yr	Maternal age $\leq$ 33 yr
HighSchool	`medu`	High-school education completed	Otherwise
FullPrenatal	`prenatal`	Adequate prenatal care	Inadequate / none
Smoker	`cig_0`	Any prenatal smoking	No smoking

Dimension Reduction

After the first step in preprocessing, the data is encoded as dichotomous indicator variables and one response column of total recorded birth-weight outcomes for the 3 million records. There are $2^7 = 128$ possible combinations for the predictors, each representing a unique *class* of maternal and infant characteristics. Aggregating observations by class greatly reduces the computational load while preserving interpretability and necessary information of features, enabling discernment of risk factors with minimal computational burden.

Converting Birth-Weight into Count Variables

The final step, transforms the dataset from 3.6 million by 237 to 128 by 11. We define a sequence of decline-quantile cut points based on 10% quantile increments, to segment the LBW region (from 0-to-2.5 kg) creating 10 total LBW categories. By using the previous year's identical dataset from 2020 in segmenting these categories, we eliminate problems with "double-dipping" and bias in later estimates. This dimension reduction drastically consolidates the dataset while providing a straightforward way to retrieve the number of observations within a given class and birth-weight category. The specific cut-point values and their prior assignment are shown in Table 2.2.

Table 2.2 the two types include: LBW and NBW, and the other restricted only to LBW, where NBW is defined as any newborn with greater than 2.5 kg at birth (Wikipedia contributors, 2025a). Once the tree is fit, they are called the "full" and "LBW-only" models respectively. The NBW observations are aggregated into one column called `counts_above_2.5kg`, and will serve as the 11th birth-weight category in the consolidated counts dataset. When given to CART, it is of primary concern how the inclusion of this column changes the tree construction, variable selection, and stability of estimates. Additionally, the prior construc-

Table 2.2: Birth-weight quantile cut points and Dirichlet priors

	LBW + Normal		**LBW only**	
Quantile	**Range (g)**	**Prior (%)**	**Range (g)**	**Prior (%)**
Q1	227–1170	0.84	227–1170	10
Q2	1170–1644	0.84	1170–1644	10
Q3	1644–1899	0.83	1644–1899	10
Q4	1899–2069	0.83	1899–2069	10
Q5	2069–2183	0.87	2069–2183	10
Q6	2183–2270	0.83	2183–2270	10
Q7	2270–2350	0.86	2270–2350	10
Q8	2350–2410	0.93	2350–2410	10
Q9	2410–2460	0.71	2410–2460	10
Q10	2460–2500	0.80	2460–2500	10
Normal	>2500	91.67		

tion is heavily skewed in the full model, where the NBW column has probability of 91.67%, while the LBW-only is a uniform 10% prior probability across all 10 LBW categories by construction.

In summary, this preprocessing consolidation yields 10 discretized quantile birth-weight categories used to allocate all observations into *counts*. This provides a detailed gradations of the LBW region, and adding the aggregated NBW category provides the full range of birth-weight outcomes in the dataset. This approach prevents scarcity in any one category, and the data will be called *counts data* from here forward.

Marginal Dirichlet-Multinomial (DM) Likelihood

Introduction

In the previous section, we established the discretized birth-weight categories for this study. Resulting in 10 LBW categories of quantile increments of 10% and the aggregated 11th category for NBW. Modeling the distribution of birth-weight counts must require handling categorical partitions of all birth-weight categories, typically with small sample sizes in observed samples, thus *zero-counts* in one or more categories. Relying on the standard maximum likelihood estimation (MLE) approach of observing raw proportion of observation counts, often results probability of zero assigned to some categories not observed in the sample. If we disregard this issue entirely, the MLE implicitly eliminates such categories that have not already been recorded so far, which is an unreasonable assumption for further inference. Thus, a smoothing technique is required to prevent such categories with zero-counts from being *impossible* in future birth-weight observations, while still balancing small enough probabilities to reflect how rarely (or never) such events appear in the data, called *overdispersion*.

One powerful and effective solution is through the Dirichlet-Multinomial (DM) model, which relies upon the Dirichlet-Multinomial conjugate pair, and interpretable via the Pólya urn scheme (Mimno, 2025; Gundersen, 2020; Minka, 2000). The Dirichlet's overdispersion, effectively injects "pseudo-counts" or "zero-inflating" prior observations in all birth-weight categories (Mimno, 2025; Wikipedia contributors, 2025c). This ensured that the marginal likelihood, or evidence, for any category remains strictly positive, i.e. non-zero probability assignment for unobserved categories (Wikipedia contributors, 2025c). Here, $\alpha = (\alpha_1, \ldots, \alpha_K)$ represent the Dirichlet hyperparameters, where each α_k functions as a prior count for its respective category. In this section, we will discuss the notation of the natality dataset, formally derive the marginal DM likelihood criterion, and discuss how it is

implemented in CART. By construction of the counts data, the actualized observations will follow a multinomial distribution, with a minor technicality in the standard form discussed in Section 2.

The DM likelihood is an appropriate choice for birth-weight modeling given count observations. The Bayesian splitting criterion gives the model flexibility, CART offers an interpretable algorithm for disentangling interactions, and modeling the full spectrum of birth weights gives the breadth for targeted LBW prediction and intervention.

Data Format

Before deriving the marginal DM likelihood, it is best to describe the data format. We have $K \in \{10, 11\}$ birth-weight categories, varying between 10 and 11 depending on model scope. The predictor matrix is fixed at $\mathbf{X} \in \{0,1\}^{N\times 7}$, where $N = 128$ is the number of total rows (and classes) where each row $\mathbf{x}_i \in \{0,1\}^7$ represents a dummy-encoded predictor vector for a given class i. The count data is the response matrix $\mathbf{Y} := [n_{i,k}]_{i=1,\ldots,N}^{k=1,\ldots,K}$ of dimensions $N \times K$, where $n_{i,k}$ is a number of birth observations for class i, and quantile category k. For class $i = 1,\ldots,N$, each $\mathbf{x}_i$ of $\mathbf{X}$ is a 7-dimensional feature vector representing a unique maternal-infant combination of predictors. The corresponding row $\mathbf{y}_i = (n_{i,1},\ldots,n_{i,K})$ in $\mathbf{Y}$ is of length K of counts observations for birth-weight categories $k = 1,\ldots,K$. In other words, for each unique predictor class $\mathbf{x}_i$, $\mathbf{y}_i$, tells us how many births fell into each category k with probability θ_i. This setup is from the preprocessing steps described in Section 2, which dramatically consolidate the dataset into counts. The N total classes each $\mathbf{x}_i, \mathbf{y}_i$ pair concisely represent all predictors and corresponding birth-weight response frequencies.

To illustrate, if $K = 3$ and a particular predictor vector $\mathbf{x}_i$ appears 10 times in the data, with outcomes of 6 in class 1, 3 in class 2, and 1 in class 3, then $\mathbf{y}_i = (6,3,1)$ and $N_i = 6+3+1 = 10$. We can apply the DM likelihood model for multivariate, multinomial counts data.

Treatment of the Multinomial Coefficient

Before deriving the marginal DM likelihood, we will clarify why the usual multinomial coefficient is omitted. After collapsing the dataset into N classes, each class i is summarized by its counts vector $\mathbf{y}_i = (n_{i,1}, \ldots n_{i,K})$ with the total counts in class i represented as $N_i = \sum_{k=1}^{K} n_{i,k}$. In the classic multinomial probability mass function, the factor

$$\text{MultinomialCoeff}(n_{i,1}, \ldots n_{i,K}) \;=\; \binom{N_i}{n_{i,1}, \ldots n_{i,K}} \;=\; \frac{N_i!}{\prod_{k=1}^{K} n_{i,k}!}$$

enumerates every possible permutations of N_i births inside of class i. Because the information of the raw sequence of individual counts is *not kept* by consolidating the dataset into counts data, we no longer model the possible orderings of N_i births. Because this coefficient is constant with respect to the category probability vector θ_i (Wikipedia contributors, 2025c), it plays no role in the likelihood-split comparisons and therefore is omitted from our criterion.

To justify why this is the case, suppose we partition N into two splits (instead of 10 or 11) N_1, N_2 where $N = N_1 + N_2$. Then the partitioned counts N_1, N_2 have less possible permutations of $n_{1,K}$ and $n_{2,K}$ counts, respectively. That is to say:

$$\binom{N}{n_1, \ldots, n_K} > \binom{N_1}{n_{1,1}, \ldots, n_{1,K}} + \binom{N_2}{n_{2,1}, \ldots, n_{2,K}}$$

Derivation

The hierarchical model structure is as follows:

$$\mathbf{x}_i = \text{ dummy-encoded predictor vector for class } i, \quad i = 1, \ldots, 128$$

$$\mathbf{y}_i \mid \theta_i, \mathbf{x}_i \sim \text{AdjustedMultinomial}\,(N_i, \theta_i)$$

$$\theta_i \sim \text{Dirichlet}\,(\alpha = (\alpha_1, \ldots, \alpha_K)) \qquad \text{(prior)}$$

$$\theta_i \mid \mathbf{y}_i, \mathbf{x}_i \sim \text{Dirichlet}\big(\alpha + \mathbf{y}_i\big) \qquad \text{(posterior)}$$

Where $N_i = \sum_{k=1}^{K} n_{i,k}$ is the row total and $\theta_i = (\theta_{i,1}, \ldots, \theta_{i,K})$ is the true (but unknown) category probabilities for class i.

From here forward, we will omit the index i from $\mathbf{x}_i, \mathbf{y}_i, \theta_i$ to avoid confusion. This changes to counts vector $(n_{i,1}, \ldots, n_{i,K})$ to $(n_1, \ldots, n_K)$, where n_1 is naturally interpreted as the first 10% quantile. Likewise let $\theta = (\theta_1, \ldots, \theta_K)$ be underlying Dirichlet prior, where θ_k is the probability of an observation falling into category k. The Dirichlet prior $p(\theta)$ is given hyperparameters α where each $\alpha_k > 0$ obtains the density:

$$p(\theta) \;=\; p(\theta \mid \alpha) \;=\; \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \;=\; \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(\alpha_0)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

for $\theta_k \geq 0$, and $\sum_k \theta_k = 1$. Here, $B(\alpha)$ is the multivariate Beta function, serving as the normalizing constant. $B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)}$, with $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ for brevity. The Dirichlet component encodes our prior belief about the probabilities of θ_k acting as "prior-counts" of category k (Wikipedia contributors, 2025c). Given θ, the probability of observing a specific count outcome follows the adjusted multinomial likelihood. This is the likelihood of the category k for a given θ_k.

$$p(\mathbf{y} \mid \theta, \mathbf{x}) \;=\; \underbrace{\frac{N!}{\prod_{k=1}^{K} n_k}}_{\text{omit}} \prod_{k=1}^{K} \theta_k^{n_k} \;=\; \prod_{k=1}^{K} \theta_k^{n_k}$$

Under this structure, the joint density of the data and latent probability vector is the product of the Dirichlet prior and adjusted multinomial likelihood. Substituting both components yields:

$$\begin{aligned} p(\mathbf{y}, \theta \mid \mathbf{x}) &= p(\mathbf{y} \mid \theta, \mathbf{x}) p(\theta) \\ &= \underbrace{\prod_{k=1}^{K} \theta_k^{n_k}}_{\text{Adjusted Multinomial } p(\mathbf{y}|\theta,\mathbf{x})} \underbrace{\frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}}_{\text{Dirichlet prior } p(\theta)} \\ &= \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{(n_k + \alpha_k) - 1} \\ &\propto \left(\theta_1^{(n_1+\alpha_1)-1}, \ldots, \theta_K^{(n_K+\alpha_K)-1}\right) \sim \text{Dirichlet}\big(\alpha + \mathbf{y}\big) \quad \text{(posterior)} \end{aligned}$$

Here we see the Dirichlet prior's effect is to "shift" of exponent in $\theta_k^{n_k}$ by $\alpha_k - 1$, adding α_k to n_k in the exponent (Mimno, 2025). To achieve the goal of the marginal likelihood, we integrate over all possible θ of the joint density.

$$\begin{aligned} p(\mathbf{y} \mid \alpha, \mathbf{x}) &= \int_{\theta} p(\mathbf{y}, \theta \mid \mathbf{x}) p(\theta) d\theta \\ &= \int_{\theta} \left(\prod_{k=1}^{K} \theta_k^{n_k} \right) \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} d\theta \\ &= \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \int_{\theta} \prod_{k=1}^{K} \theta_k^{(n_k + \alpha_k) - 1} d\theta \end{aligned} \tag{2.1}$$

The constant $\frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)}$ can be pulled outside of the integral since these do not depend on θ. The integral in the final line is recognizable as the normalization integral of a Dirichlet distribution, the Beta function with parameters $(\alpha_1 + n_1, \alpha_2 + n_2, \ldots, \alpha_K + n_K)$ can simplify this line. Define $m_k = \alpha_k + n_k$

$$
\begin{aligned}
\int_\theta \prod_{k=1}^{K} \theta_k^{m_k-1} d\theta \;&=\; B(\mathbf{m}) \\
&= \frac{\prod_{k=1}^{K} \Gamma(m_k)}{\Gamma(\sum_{i=1}^{K} m_k)} \\
&= \frac{\prod_{k=1}^{K} \Gamma(n_k+\alpha_k)}{\Gamma(\sum_{i=1}^{K} n_k+\alpha_k)} \\
&= \frac{\prod_{k=1}^{K} \Gamma(n_k+\alpha_k)}{\Gamma(N+\alpha_0)}
\end{aligned}
$$

where $N = \sum_{k=1}^{K} n_k$. Now substitute the last line back into Equation 2.1 to achieve the final closed-form marginal DM likelihood (Wikipedia contributors, 2025c):

$$
\begin{aligned}
p(\mathbf{y} \mid \alpha, \mathbf{x}) \;&=\; \frac{1}{B(\alpha)} B(\alpha+\mathbf{y}) \\
&= \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(n_k+\alpha_k)}{\Gamma(N+\alpha_0)} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(N+\alpha_0)} \prod_{k=1}^{K} \frac{\Gamma(n_k+\alpha_k)}{\Gamma(\alpha_k)}
\end{aligned}
\tag{2.2}
$$

Taking the log transform yields the equivalent form which we directly implement Equation 2.3 in the objective function criterion in `rpart`:

$$
\begin{aligned}
\log p(\mathbf{y} \mid \alpha, \mathbf{x}) \;&=\; \log\Gamma(\alpha_0) \;-\; \log\Gamma(N+\alpha_0) \\
&\quad + \sum_{k=1}^{K} \Big(\log\Gamma(n_k+\alpha_k) \;-\; \log\Gamma(\alpha_k) \Big).
\end{aligned}
\tag{2.3}
$$

The equations above can be recognized as the DM distribution, sometimes called Compound Multinomial or Pólya's urn distribution (Mimno, 2025). This can be broken down component-wise to give an intuitive understanding: $\frac{\Gamma(n_k+\alpha_k)}{\Gamma(\alpha_k)}$ shows observing n_k instances of category k updates the prior count α_k, the ratio $\frac{\Gamma(\alpha_0)}{\Gamma(N+\alpha_0)}$ ensures that all the probabilities for categories sum to 1, yielding the normalization factor across joint observation counts (Gundersen, 2020; Wikipedia contributors, 2025c). While $\alpha_k > 0$, the likelihood will be

nonzero even if $n_k = 0$ for some categories and the prior α_k acts as the smoothing term guaranteeing $p(\mathbf{y} \mid \alpha) > 0$ for all possible outcomes (Mimno, 2025).

Splitting with Adjusted DM Likelihood in CART

Equation 2.3 is the final log marginal DM likelihood that will be implemented in CART. Now we can understand how CART uses this criterion for evaluating possible splits of the data.

A successful decision tree will try to rid as much variation, or impurity, within a subgroup as it can, by proposing and evaluating splits on the predictors. For any dichotomous predictor under consideration, CART proposes a left and right split on this predictor and chooses the split that better explains the data. The split that is chosen is the one with the highest likelihood, indicating a better model fit. Maximizing the improvement gain is equivalent to minimizing the node impurity.

Using the adjusted marginal log-likelihood as our impurity measure, denote $\mathscr{L}_{\text{node}}$ for the likelihood for a node's data. Let a parent node count observations, $\mathbf{y}_{\text{parent}}$ be split into $\mathbf{y}_{\text{left}}, \mathbf{y}_{\text{right}}$ we calculate first: $\mathscr{L}_{\text{parent}} = \log p(\mathbf{y}_{\text{parent}} \mid \alpha, \mathbf{x})$ then, $\mathscr{L}_{\text{left}} = \log p(\mathbf{y}_{\text{left}} \mid \alpha, \mathbf{x})$ and $\mathscr{L}_{\text{right}} = \log p(\mathbf{y}_{\text{right}} \mid \alpha, \mathbf{x})$ and calculate the improvement gain of the split as:

$$\Delta\mathscr{L} = \mathscr{L}_{\text{left}} + \mathscr{L}_{\text{right}} - \mathscr{L}_{\text{parent}}$$

Essentially, if $\Delta\mathscr{L}$ is positive, the split results in an *improvement* where the DM criterion favors the split with more homogeneity. If the multinomial coefficient was included then for any partition, $\mathscr{L}$ would directly reward partitions with the larger number of possible orderings of the data. As noted in Section 2, N would be "better" than the two smaller groups of size N_1 and N_2, irrespective of how the counts are distributed. The omission of the coefficient makes the adjusted log-likelihood evaluate splits solely on how they reflect the distributional fit.

To illustrate how CART evaluates $\Delta\mathscr{L}$, consider the scenario where a parent node with N observations is evenly split (50/50) between two nodes, with a symmetric prior $\alpha_1 = \alpha_2$. Now we evaluate some predictor to split on. The marginal likelihood of both child nodes might not exceed that of the parent since its split evenly thus no split is chosen. Conversely, if there's a predictor where one node gets all LBW observations, and the other all NBW, the combined likelihood of child nodes would be much higher than the parent. This would cause a large $\Delta\mathscr{L}$ and be a significant split for CART. This matches our intuitive goal of wanting to split on improved subclassification of the population.

Chapter 3

TREE-BASED NONPARAMETRIC BIRTH-WEIGHT MODELING

Introduction

We model birth-weight risk using a tree-based, nonparametric approach while still modeling the full spectrum of birth-weight outcomes, we focus primarily on gradients of increased LBW risk among possible predictors. Birth-weights are first grouped into quantile-defined categories so the tree can detect subtle shifts in risk across the finer LBW region, providing a more granular view. Because all factors potentially effect birth weight, our goal is to pinpoint combinations of predictors that consistently mark subpopulations with an elevated LBW incidence.

To avoid "double dipping," or using the same observations to both fit the model and set its prior, we derive LBW quantile cut points and subsequently construct the Dirichlet hyperparameters from the *previous* year's data (2020) and apply them unchanged to the 2021 dataset. Figure 3.1 shows the resulting prior vector α. The left-panel shows full model's expected proportion of observations in both NBW and LBW categories, and the right-panel shows the LBW-only model's uniform prior due to the 10 LBW deciles. Year-to-year, the birth-weight distributions are remarkably stable at the national level, making 2020 a suitable proxy for 2021.

Seven dichotomous maternal-infant predictors maternal race (`mrace15`), smoking during pregnancy (`cig_0`), marital status (`dmar`), maternal age (`mager`), education (`meduc`), adequacy of prenatal care (`precare5`), and infant gender (`sex`) are given to CART, searching recursively for the largest improvement gain. To confirm stability and reliability of the split predictors we will create a 10,000 bootstrap ensemble and compare the variable

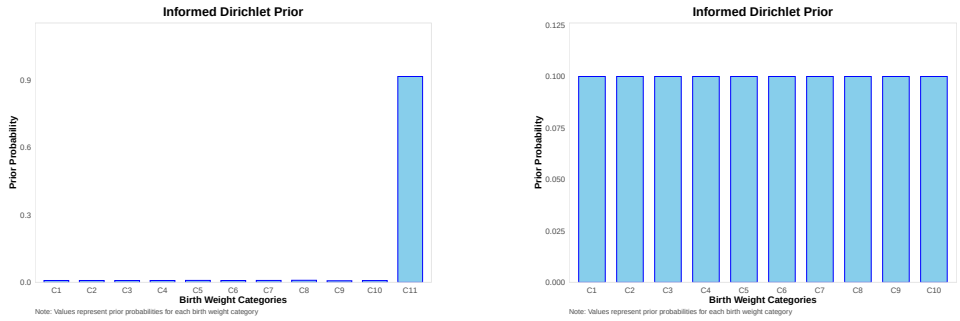
selection and stability across the trees. This will be further elaborated on in Section 3.

Constructing the Informed Prior

Quantile Cut Points & Informed Prior from 2020 Data

All birth weights less than or equal to the threshold of 2.5 kg are divided into ten deciles comprising the 10% quantiles mentioned earlier. The LBW-region then smoothly transitions from "extremely low" (lowest 10%) to "moderately low" (highest 10%), and pool all NBW observations into the eleventh category denoted as `counts_above_2.5kg`. The large NBW group in the full model allows us to see how this dominant category influences tree splits and variable selection. Under different model scopes, the full and LBW-only model contrasts the focus of the models. The LBW-only model centers its attention around variation *within* the LBW-region in Section 3.

From the 2020 proportions we construct the Dirichlet prior, $\alpha = (\alpha_1, \ldots, \alpha_K)$. In Figure 3.1 the priors show the strength in NBW under the full model and the uniform prior under the LBW-only models across all categories.



(a) Full model informed prior

(b) LBW-only model informed prior

Figure 3.1: Comparison of informed Dirichlet priors based on 2020 quantiles

DM-CART

Custom Objective Function & Splitting

We fitted CART using `rpart`, in R (Therneau *et al.*, 2025), replacing the default Gini index with the adjusted marginal log DM likelihood or simply the log likelihood for brevity. Serving as our objective function, it output scores based on the reduction in deviance of a given split, i.e. negative log likelihood output (Therneau *et al.*, 2025; Therneau and Atkinson, 2023). When splitting, `rpart` computes the left and right split deviance calculation as in Section 2. For instance, suppose we propose a split on smoking status. The objective function evaluates separating the data into smoking and non-smoking mothers, as "distinct" on the count response vector. The improvement is the reduction in deviance, rewarding meaningful distributional shifts, thereby reducing heterogeneity in the node. Thus, distinctness here means the improvement gain from splitting under the predictor in question. Throughout, “risk” is used heuristically to describe a subpopulations relative prevalence of LBW outcomes.

Tree Results and Insights: Full Model

For the full model, Figure 3.2 shows the hierarchical structure, and Figure 3.6 ranks improvement gain for each split. The full model splits as follows:

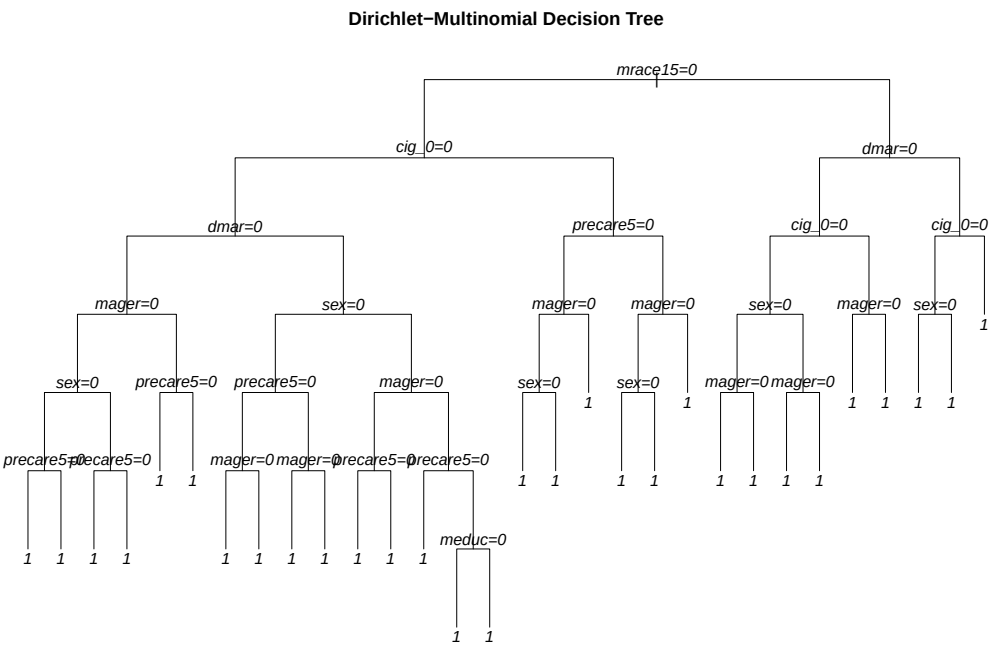


Figure 3.2: Full model tree structure

The root node contains all of the 2021 birth counts, and naturally the vast majority fall above 2.5 kg. The first, and largest deviance reduction, comes from separating the counts based on race, i.e. Black mothers (`mrace15=1`) from all other mothers (`mrace15=0`). This split yields the largest difference in birth-weight profiles and is shown to be the most informative predictor. This split aligns with documented discrepancies of race, playing a critical role in LBW incidence (Kaiser Family Foundation, 2025; Colen *et al.*, 2006). Among Black mothers, smoking status supplies the next greatest improvement, whereas among non-Black mothers, smoking status is considered *only after* marital status. That is to say, smoking most strongly differentiates outcomes for Black mothers, while partnership status matters more for non-Black mothers. These results show that further splits occur based on racial demographics. Further, analysis demonstrate that the overall highest risk subpopu-

lations are among unmarried Black smokers, where $\texttt{mrace15} = 1, \texttt{cig\_0} = 1, \texttt{dmar} = 0$. Overall, race, smoking, and marital status jointly account for the bulk of the total improvement, confirming earlier evidence that Black smokers and unmarried non-Black mothers constitute the highest-risk subpopulations among racial demographics (Kaiser Family Foundation, 2025; Colen *et al.*, 2006; Delcroix-Gomez *et al.*, 2022) and Figure 3.2 and Figure 3.6 visualize these results. Further splitting down the branch of Black smokers provides further nuance of the highest-risk subgroups ($\texttt{mrace15} = 1, \texttt{cig\_0} = 1, \texttt{dmar} = 0$). This node splits on infant gender, where on average, female infants weigh less than males (Van Vliet *et al.*, 2009).

Generally, the lowest-risk subpopulations are where mrace15=0 — recall that smoking status is considered after marital status of the mother. Despite the ordering of ranked splits, smoking is a direct determinant of LBW incidence (Delcroix-Gomez *et al.*, 2022). After socioeconomic and demographic variables are considered, the model emphasizes more biological and genetic predictors, namely gender, prenatal care, and maternal age. The lowest-risk groups where $\texttt{mrace15} = 0, \texttt{dmar} = 1, \texttt{cig\_0} = 0$, with age being the final predictor considered. Moreover, this branch has a depth of 4 while the highest-risk subgroups have a depth of 6 and 7.

Surprisingly, the only case where education status (meduc) was used was when the mother was Black, non-smoker, below 33 years old, without adequate prenatal care, and had a female newborn, (i.e. $\texttt{mrace15} = 1, \texttt{cig\_0} = 0, \texttt{dmar} = 0, \texttt{sex} = 0, \texttt{mager} = 0, \texttt{precare5} = 0$). Given that education has been noted by many (Martinson and Choi, 2019) to have an effect on socioeconomic conditions, namely earnings, which might not be as strong of a predictor as initially thought.

The full model structure provides insights into the direct risk stratification. This approach finds race, smoking status, and marital status as the dominant predictors, ranked as the top three splits in Figure 3.6. The model shifts then toward biological predictors of

maternal age, infant gender, and adequacy of prenatal care. Lastly education status is only used in one split, providing the least improvement.

Tree Results and Insights: LBW-Only

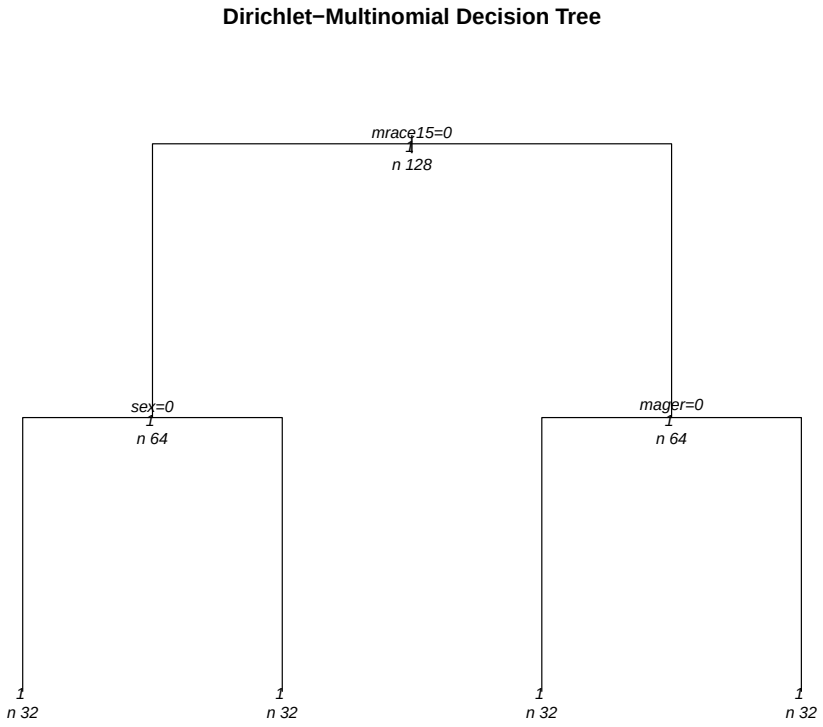


Figure 3.3: LBW-only model tree structure

As seen in Figure 3.3, the restriction to only LBW observations drastically changes the tree's structure. The distributional contrast between NBW and LBW disappears by this restriction. Initially, this model has more homogeneity in the data yielding less drastic improvements. Race again dominates the root split, with the second-level splitters being now infant gender for the `mrace15=1` branch, and maternal age for the `mrace15=0` branch; with

smoking status, and marital status *never* appearing. By every infant already being below 2.5 kg., behavioral and socioeconomic factors that distinguish NBW and LBW, no longer provide useful partitions. Instead, biological factors explain within-LBW heterogeneity. A key takeaway from contrasting the two models is that the LBW-only model suggests Black mothers *continue* to have a higher incidence of LBW newborns.

Variable importance among the full and LBW-only model are drastically different as well. Figures 3.6 and 3.7 order predictors by the sum of total deviance each eliminates across all splits, showing the contrast between predictor usage. The barplots reflects the greedy search of CART, where early splits absorb a large share of heterogeneity and later splits improve the fit only marginally, regardless of their actual effect on the response (Centre for Speech Technology Research, nd).

Depth-Controlled Model Comparison

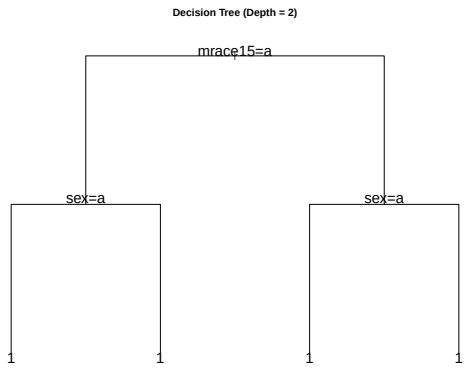
To investigate how different models select variables as the tree grows, we refit both the full and LBW-only CART models at maximum depths of 2, 3, 4, 5, while explicitly tracking the smoking predictor to discern its roles among other predictors in different modeling contexts. Figures 3.4 and 3.5 visualize the trees.

As the depths increase for the full model (Figure 3.4), the number of terminal nodes expands from 4 at depths 2, to 19 at depth of 5, while the number of predictors rises from 3 to 6. Race, infant gender, and marital status appear in every depth while maternal age and prenatal care are included at depth 4. Moreover, `cig_0` is selected at only at depths 4 and 5, suggesting that once additional socioeconomic and biological variables are available, smoking contributes little deviance reduction with the full spectrum of birthweight outcomes.

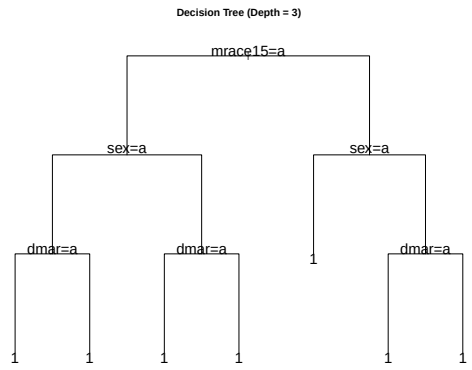
The LBW-restricted trees (Figure 3.5) the growth complexity stagnates after depth of 3, starting at 4 terminal nodes at depth of 2 growing to only 5 at depth 3 through 5.

The lessened number of leaves reflects the large homogeneity in the restricted LBW data. Like the full model, the initial splitter is race and subsequent splits consider infant gender, maternal age, then considering prenatal care adequacy *only for Black male newborns* ($\texttt{mrace15} = 1, \texttt{sex} = 1$). These results highlight the stark contrast between the full model's ending complexity. Here, we consider how race, maternal age, infant gender, and prenatal care effect LBW severity among LBW cases. Crucially, the predictors of `cig_0`, `dmar`, and `meduc` are never considered, demonstrating that such socioeconomic factors do not play a critical role in identifying added risk among LBW outcomes. That is to say that, smoking, marital status, and educational attainment do not significantly contribute to increased LBW severity.

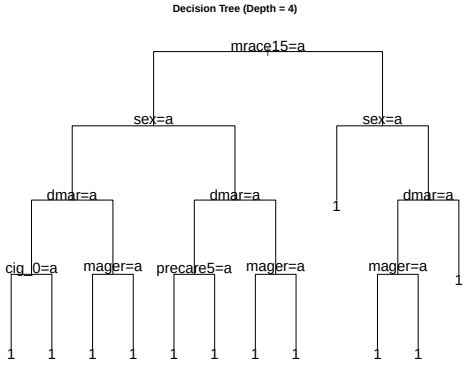
It is clear that the predictor hierarchy and prioritization has shifted when the depth parameter is restricted compared to the first fit LBW-only model in Section 3. Due to this difference, we will employ a two-tier bootstrap procedure to confirm stability of variable selection and importance.



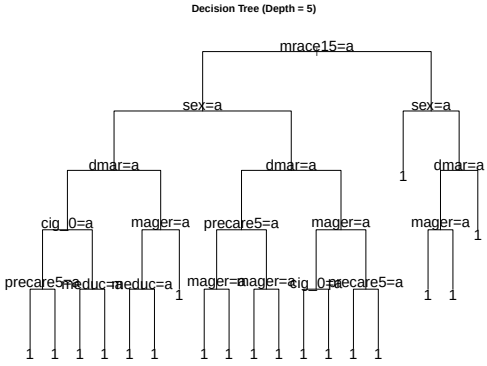
Maximum depth = 2



Maximum depth = 3

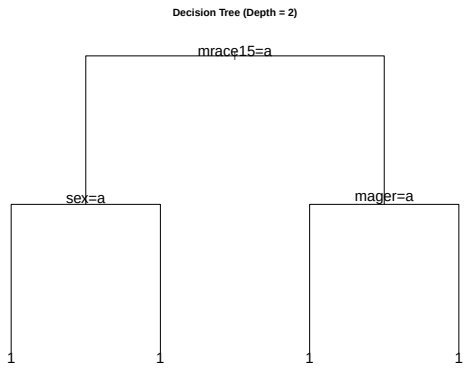


Maximum depth = 4

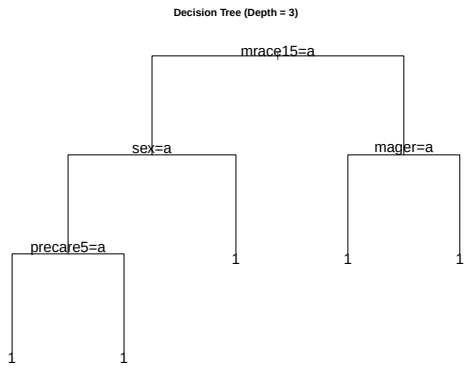


Maximum depth = 5

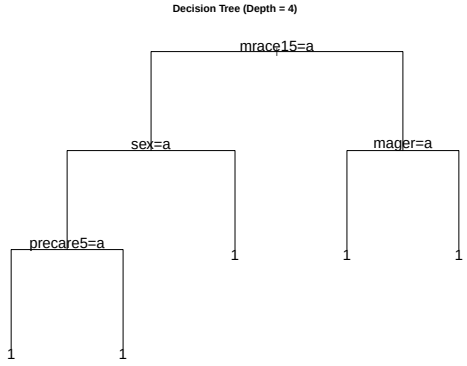
Figure 3.4: Full model tree structures, showing growth patterns at different maximum depths (2,3,4,5). The trees demonstrate variable selection patterns with increasing depth, highlighting the growing complexity of the model structure.



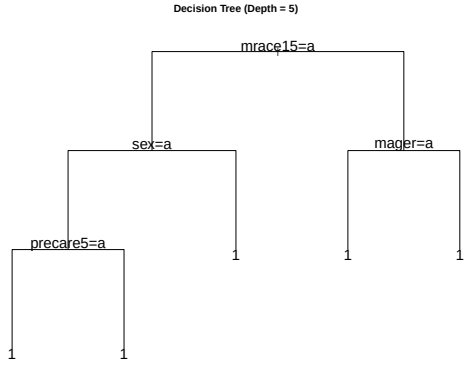
Maximum depth = 2



Maximum depth = 3



Maximum depth = 4



Maximum depth = 5

Figure 3.5: LBW-only model tree structures, showing growth patterns at different maximum depths (2,3,4,5). The trees demonstrate variable selection patterns with increasing depth, highlighting the limited growth among LBW outcomes.

Bootstrap Analysis & Methodology

Introduction

To test whether the splits observed in Section 3 are specific to one realization of the 2021 data, we construct an ensemble of $B = 10,000$ parametric bootstrap trees. For each replicate, resampling perturbs the data into two levels, mirroring the sampling hierarchy in the DM model. That is to say, we will focus on (1) *between-class* counts and (2) *within-class* counts. These steps will be the stages, or tiers, of the bootstrap procedure. First, we randomize how the total number of births T is partitioned across the $N = 128$ predictor classes and secondly, given the total class partitions, we randomize how those births are allocated among the K birth-weight categories. This will deliver uncertainty estimates that are coherent with our criterion used to fit each tree. Throughout, "row" and "class" are synonymous.

The goal of the bootstrap procedure is to provide robust and stable probability estimates $\hat{\pi}_{i,k}$ for each category k. The frequencies in Table 3.1 therefore have a defensible interpretation as the bootstrap probabilities of variable inclusion.

Justification of the Two-Tier Bootstrap Procedure

To motivate the multinomial assumption in each tier, consider the consolidated counts data. It is an *aggregated* $N \times K$ matrix where the cell entries $n_{i,k}$ are sums of total number of births for class i and category k. These sums are not *individual* observations. Treating the row vector $\mathbf{y}_i$ as an i.i.d. "case" would breech the key independence assumption that underpins case-resampling (Davison and Hinkley, 2021, slide 47) (Hrba *et al.*, 2022). Moreover, the counts data holds the same structure as a contingency table: conveying the frequencies of any two multivariate vectors, in this case class profiles by birth-weight categories (Wikipedia contributors, 2025b). De-aggregating these frequencies into individual

observations before performing bootstrap resampling is required to preserve the within-row dependence, preserving the data structure (stats.stackexchange.com, 2025a), precisely because each row is a vector of summary statistics (i.e. sum of total row observations) (Wikipedia contributors, 2025e). Further, when sampling any row, the counts vector has fixed proportions of LBW and NBW counts, thereby providing no within-row randomness. Since NBW proportion greatly exceeds that of the LBW counts this would greatly overshadow the LBW variability.

When individual birth records cannot be recovered, the solution is a model-based *parametric* bootstrap that respects both levels of randomness while maintaining faithful to the DM model introduced in Section 3.

Methodology & Procedure

Formally, the two-tier bootstrap resampling procedure is defined here. First, T defines the grand total counts. $\mathbf{p} = (p_1, \ldots, p_N)$ are the empirical proportions across N classes. The *bootstrap probability estimates* for class i across K categories are represented as, $\hat{\pi}_i = (\hat{\pi}_{i,1}, \ldots, \hat{\pi}_{i,K})$, and are the mean across the B bootstrap replicate trees. Additionally, the predictor matrix $\mathbf{X}$ remains fixed; only response counts are resampled.

$$
\begin{aligned}
T &= \sum_{i=1}^{N} \sum_{k=1}^{K} n_{ik} && \text{(grand total of births)},\\
N_i &= \sum_{k=1}^{K} n_{ik} && \text{(row total for class } i\text{)},\\
p_i &= \frac{N_i}{T}, \qquad \mathbf{p} = (p_1, \ldots, p_N) && \text{(empirical class shares)},\\
\hat{\pi}_i &= (\hat{\pi}_{i,1}, \ldots, \hat{\pi}_{i,K}) && \text{(posterior DM mean for class } i\text{)}.
\end{aligned}
$$

Recall that the posterior distribution in Equation 2.1 for class i is:

$$
\theta_i \mid \mathbf{y}_i, \mathbf{x}_i \sim \text{Dirichlet}\big(\alpha + \mathbf{y}_i\big), \tag{2.1}
$$

whose mean is:

$$\hat{\pi}_{i,k} \;=\; \mathbb{E}\big[\theta_{i,k} \mid \mathbf{y}_i\big] = \frac{n_{ik} + \alpha_k}{N_i + \alpha_0}, \qquad \alpha_0 = \sum_{k=1}^{K} \alpha_k.$$

All $\hat{\pi}_{i,k}$ represent the mean bootstrap probability estimates (across B) for combination i,k. Importantly, these K estimates are fixed. Instead of directly inferring about $\theta_{i,k}$ under the posterior, we use the shrinkage estimate $\hat{\pi}_{i,k}$. This is referred to as fixed under the multinomial distribution in Tier 3 Wikipedia contributors (2025d); Duke University (nd).

Tier 1: Between-class counts resampling First we draw a new vector of class totals from a multinomial distribution, sampling once per class. This tier propagates sampling noise in relative prevalence of the N predictor profiles.

$$\mathbf{n}_i^* = (n_1^*, \ldots, n_N^*) \sim \text{Multinomial}(T, \mathbf{p}) \tag{3.1}$$

Tier 2: Within-class counts resampling Conditioned on the newly drawn total $n_i^* > 0$, resample the K category counts.

$$\tilde{\mathbf{y}}_i = (\tilde{n}_{i,1}, \ldots, \tilde{n}_{i,K}) \sim \text{Multinomial}(n_i^*, \hat{\pi}_i) \tag{3.2}$$

Since $\hat{\pi}_i = (\hat{\pi}_{i,1}, \ldots, \hat{\pi}_{i,K})$ is plugged-in and held fixed here, each vector $\hat{\pi}_i$ under class i is referred to as the bootstrap probability estimates for K birth-weight categories. Each bootstrap replicate inherits prior information via α, allowing both inter- and intra-class sampling variability to be propagated. The estimates are interpreted naturally as the best guess for the probability of a future birth from class i to fall into category k.

After the resampled counts are obtained, $\tilde{\mathbf{Y}} = [\tilde{n}_{i,k}]$ is paired with $\mathbf{X}$ and fitted with the DM–CART procedure. From each tree we record the predictor set used in splitting and the root split predictor. Aggregating across B bootstrap trees yields the frequencies reported in Table 3.1 and the predictor depth comparison in Figures 3.11, 3.12. By tier 1 capturing

the uncertainty in class prevalence, and tier 2 capturing uncertainty of the proportions of LBW and NBW within each class, these frequencies can be interpreted as the bootstrap probability of variable inclusion under the DM hierarchy.

For each class i, we compute the mean bootstrap probability estimates in $\hat{\pi}_i$ and hold this vector as fixed when we generate the replicate counts $\tilde{\mathbf{y}}_i \sim \text{Multinomial}(n^*, \hat{\pi}_i)$. Rather than resampling a new $\theta_i^{(b)} \sim \text{Dirichlet}(\alpha + \mathbf{y}_i)$ inside every bootstrap replicate b, we use the mean estimates and keep the resampling procedure focused on sampling variability of the observed counts data. This approach eliminates the need for a Monte-Carlo procedure and prior information α from being counted twice.

Variable Selection Frequency

Table 3.1 distills the frequency of variable selection across the bootstrap ensemble. These are the probability estimates for variable inclusion under each model, highlighting the relative importance under each context. In both the full model and LBW-only model, maternal race consistently emerges as the dominant predictor, appearing as the initial split variable in 100% of the bootstrap trees. This finding reinforces the conclusion drawn in earlier analyses (see Section 3) that racial disparities represent the most prominent signal in the data. Figures 3.11 and 3.12 show the distribution of each predictor's depth for the full and LBW-only model respectively.

Beyond race, the variable selection patterns diverge considerably between the two models. In the full model, six of seven variables (infant gender, marital status, prenatal care, smoking status, and race) are selected in every tree (100%), while education status is selected in approximately 37.94% of the ensemble. Despite its relatively lower selection frequency, maternal education is deeply positioned in the trees, with an average depth of 5.52 in Figure 3.11, suggesting weak but possibly contextually relevant role in specific subpopulations. In contrast, marital status and smoking status appear much closer to the root

at depths 1.69 and 1.52 respectively, indicating stronger global influence across the data.

In the LBW-only model, the frequency and depth of variable inclusion reflects the shift of model focus. While race and infant gender remain universally selected (100%), only maternal age maintains high inclusion at 97.41%, and prenatal care follows at 63.51%. The remaining variables occur rarely or not at all with marital status (9.15%), smoking status (1.96%), and maternal education (0%). The stark drop in inclusion frequency suggests that given the LBW outcomes, the model reduces its reliance on broader social determinants like education and marital status, and concentrates on variables more directly related to biological and perinatal features such as age and care access.

This interpretation is further supported by the depth analysis in Figures 3.11, 3.12. For LBW-only model, maternal race is the top splitter, followed by infant sex (mean depth of 1.12), maternal age (1.15), indicating the early and consistent splits. Prenatal care appears at an intermediate depth (2.02) and less frequently included variables exhibit greater depth, such as marital status and smoking status (2.83 and 2.94, respectively). Notably, maternal education with negligible frequency and high depth (3.00), emphasizing minimal contribution in LBW context.

Ensemble Predictions & Uncertainty

Following the bootstrap resampling procedure, each replicate yields a vector of predicted probabilities in $\tilde{\mathbf{Y}}$ for the birth-weight categories aggregated across B. For every terminal node subgroup, we take the mean across all replicates to obtain $\hat{\pi}_{i,k}$: the estimated probability that a birth in class i falls into birth-weight category k. Sampling variability is summarized by empirical 2.5- and 97.5-percentiles of each birth-weight category distribution. These 95% percentile intervals are shown along side point estimates in Tables 3.2 and 3.3. Because the interval is simply the middle 95% of the resampled values, it is distribution-free (Penn State University, nd). For some i,k combination, the interval is

read as 2.5% (or 97.5%) of replicates assign smaller (larger) probability than the reported limit (Penn State University, nd). Figures 3.8 and 3.10 display the full distribution of two contrasting maternal profiles. Such profiles are referred to as "high-risk" (class 69) and "low-risk" (class 28) and reflect reasonably assumed to face adverse and favorable birth-weight outcomes, respectively.

- *High-risk* ($i = 69$): unmarried, Black, smoking mothers under 33 with $<$ High-School education, inadequate prenatal care, delivering female infants (`mrace15` = 1, `dmar` = 0, `cig_0` = 1, `sex` = 0, `mager` = 0, `prenatal` = 0, `meduc` = 0).

- *Low-risk* ($i = 28$): married, non-Black, non-smoking mothers aged 33+, $\geq$ High-School education, adequate prenatal care, delivering male infants (`mrace15` = 0, `dmar` = 1, `cig_0` = 0, `sex` = 1, `mager` = 1, `prenatal` = 1, `meduc` = 1).

In the full model, Figures 3.8, 3.9 and Table 3.2. Specifically, in Figure 3.8 we see that both profiles have a very high predicted probability of delivering a NBW infant, yet this high-risk's NBW chance (83.6%) is roughly 7% points lower than the low-risk profile (90.9%). In Figure 3.9 we focus on the probabilities within LBW-region under the full model. This diagram shows the drastic differences of risk among the high- and low-risk profiles, with an average probability of 1.64% versus 0.908%, respectively. For the most severe LBW category (C1 or $k = 1$), the highest probability is triple that of the low-risk subgroup (1.9% vs. 0.8%). The percentile intervals are extremely narrow ($\approx \pm 0.002$) indicating remarkable stability across all bootstrap replicates.

Likewise, the LBW-only model in Figure 3.10 and Table 3.3, provides more nuance among the LBW region. The high-risk profile retains a clear disadvantage in the extreme left-tail (12.3% vs. 8.6% in C1), but the two subgroups converge in the intermediate categories, and in a few moderate LBW categories the low-risk profile even slightly exceeding high-risk subgroup (such as in C8 or $k = 8$ of Table 3.3). Percentile interval widths still

remain narrow ($\approx \pm 0.008$), illustrating that these nuanced differences are nonetheless estimated with high stability and precision.

In identifying determinants of LBW outcomes, this procedure clearly delineates a consistent and statistically reliable separation between high- and low-risk profiles, even when the absolute differences in NBW probabilities appears modest.

Table 3.1: Comparison of Variable Importance in Full Model and LBW-Only Model

	Full Model	LBW-Only Model
Initial Split Variable		
mrace15	1.0000	1.0000
Variable Frequency		
sex	1.0000	1.0000
dmar	1.0000	0.0915
mrace15	1.0000	1.0000
mager	1.0000	0.9741
precare5	1.0000	0.6351
cig_0	1.0000	0.0196
meduc	0.3794	0.0000

Note: The table shows variable frequency (proportion of trees containing each variable) and initial split variable (normalized measure of predictive contribution) for both models.

Table 3.2: Mean probability estimates $\hat{\pi}_{i,k}$ (with 95% bootstrap percentile intervals) for high- and low-risk birth-weight subgroups under the *full* model.

Category k	High-risk			Low-risk		
	$\hat{\pi}$	2.5%	97.5%	$\hat{\pi}$	2.5%	97.5%
1	1.90 %	0.0180	0.0202	0.80 %	0.0075	0.0088
2	1.72 %	0.0163	0.0181	0.86 %	0.0080	0.0096
3	1.58 %	0.0150	0.0166	0.88 %	0.0080	0.0100
4	1.63 %	0.0156	0.0170	0.90 %	0.0083	0.0102
5	1.69 %	0.0162	0.0176	0.98 %	0.0090	0.0112
6	1.61 %	0.0154	0.0168	0.91 %	0.0087	0.0098
7	1.63 %	0.0157	0.0170	0.93 %	0.0088	0.0101
8	1.78 %	0.0172	0.0185	1.08 %	0.0102	0.0117
9	1.32 %	0.0126	0.0138	0.81 %	0.0075	0.0090
10	1.52 %	0.0146	0.0158	0.96 %	0.0089	0.0106
11	83.62 %	0.8330	0.8391	90.92 %	0.9026	0.9132

Note: $\hat{\pi}$ values are reported as percentages; percentile-limit columns remain on the $[0, 1]$ scale. Estimates are based on B bootstrap replicates.

Table 3.3: Mean probability estimates $\hat{\pi}_{i,k}$ (with 95% bootstrap percentile intervals) for high- and low-risk birth-weight subgroups under the *LBW-only* model.

Category k	High-risk			Low-risk		
	$\hat{\pi}$	2.5%	97.5%	$\hat{\pi}$	2.5%	97.5%
1	12.28 %	0.1143	0.1267	8.63 %	0.0802	0.0928
2	10.61 %	0.1014	0.1090	9.73 %	0.0911	0.1030
3	9.75 %	0.0935	0.1003	9.87 %	0.0930	0.1041
4	9.90 %	0.0965	0.1018	10.18 %	0.0974	0.1064
5	10.43 %	0.1014	0.1069	10.96 %	0.1054	0.1145
6	9.74 %	0.0948	0.1011	10.16 %	0.0970	0.1050
7	9.83 %	0.0953	0.1033	10.30 %	0.0973	0.1082
8	10.57 %	0.1031	0.1088	11.42 %	0.1078	0.1218
9	8.01 %	0.0776	0.0837	8.81 %	0.0835	0.0933
10	8.87 %	0.0858	0.0941	9.94 %	0.0940	0.1053

Note: $\hat{\pi}$ values are reported as percentages; percentile-limit columns remain on the $[0, 1]$ scale. Estimates are based on B bootstrap replicates.

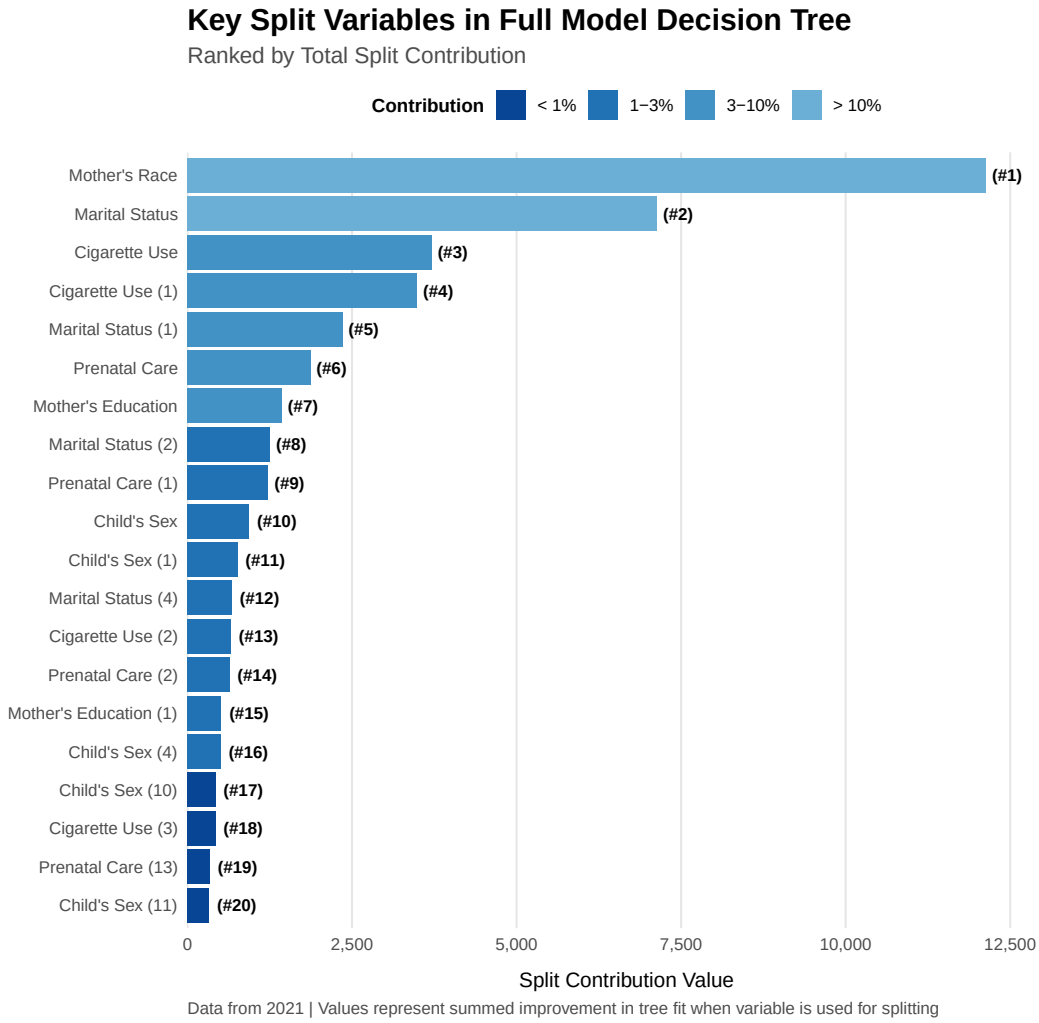


Figure 3.6: Full Model Ranked Improvement. Rankings represent summed reduction in deviance (improvement in model fit) across all nodes where each variable is used for splitting in the tree. Plot only displays top 20 ranked variables.

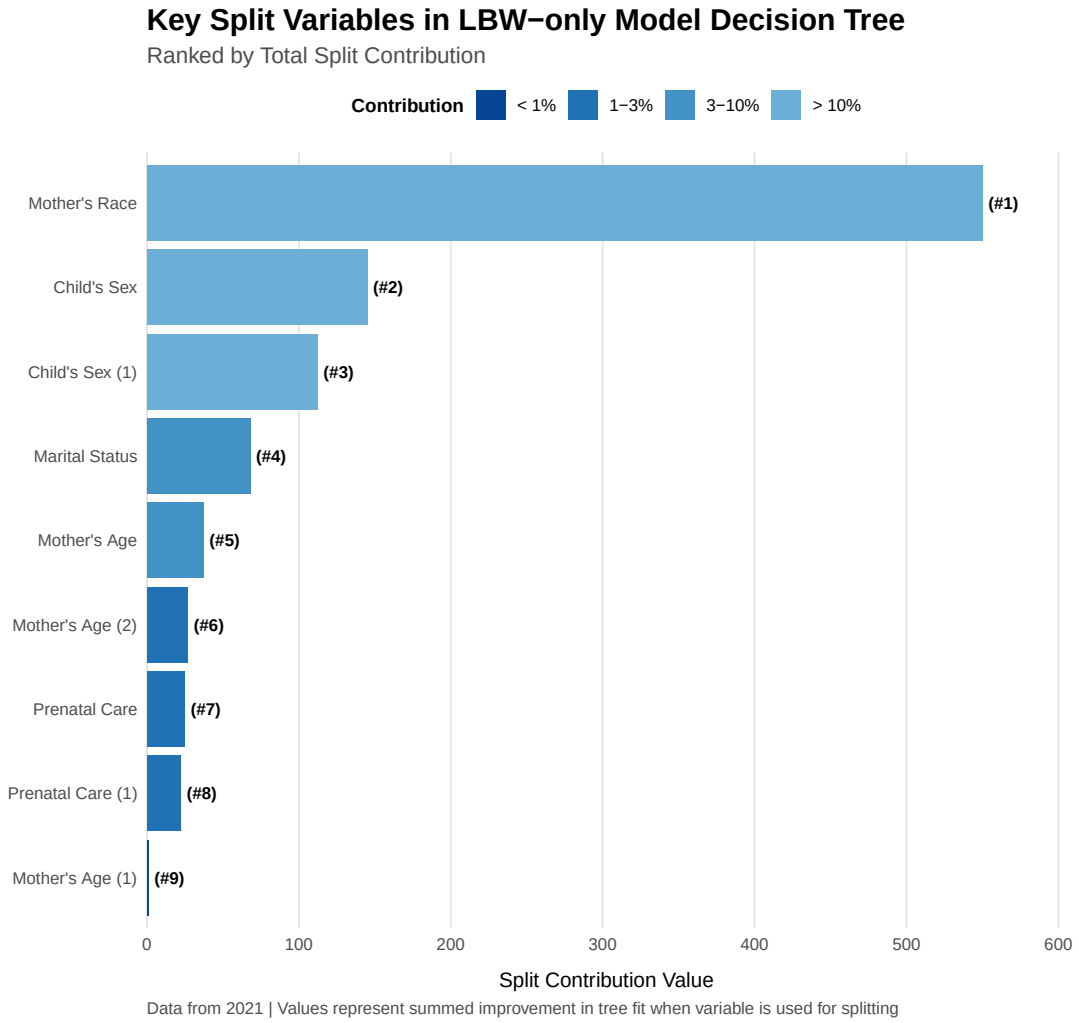


Figure 3.7: LBW-Only Model Ranked Improvement. Rankings represent summed reduction in deviance (improvement in model fit) across all nodes where each variable is used for splitting in the tree. Plot displays all ranked variables.

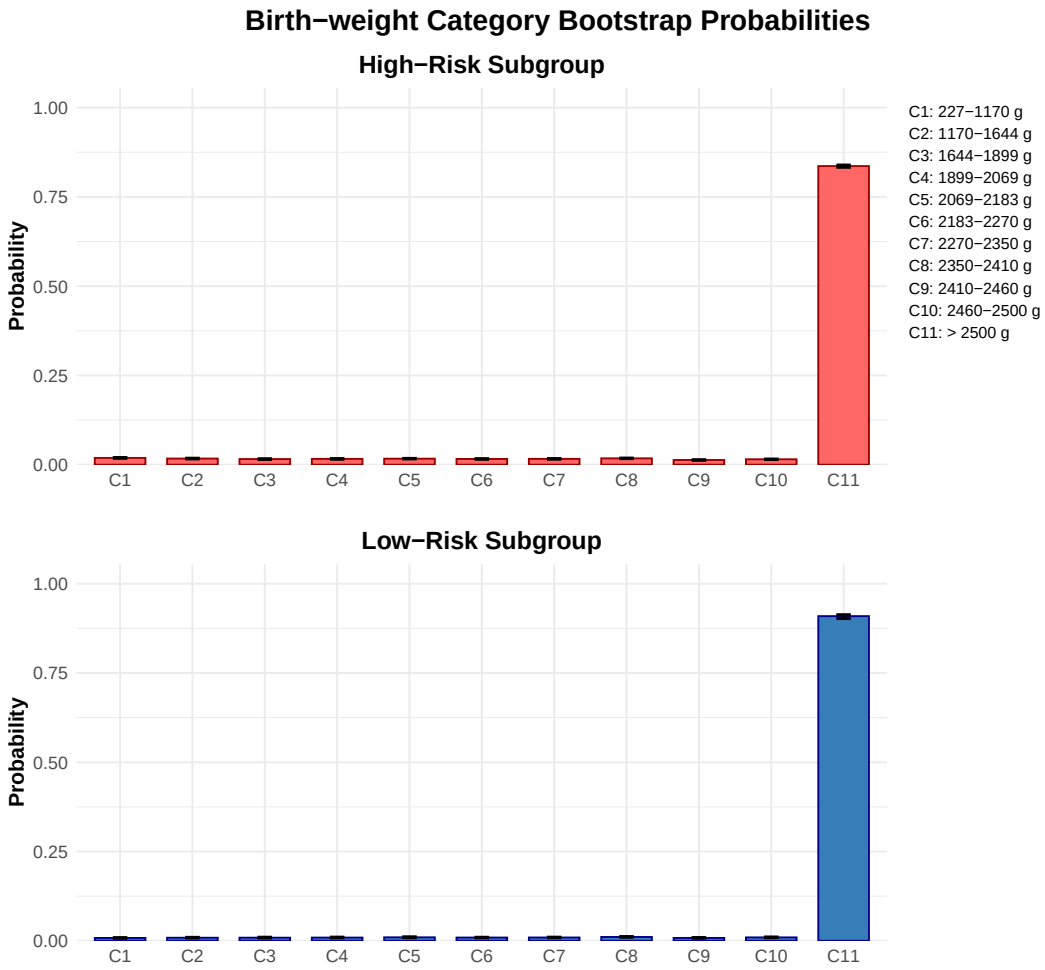


Figure 3.8: Full model: Mean bootstrap probability for each birth-weight category (C1–C11) in the high- and low-risk subgroups. Error bars show the 95 % percentile interval over the B bootstrap trees.

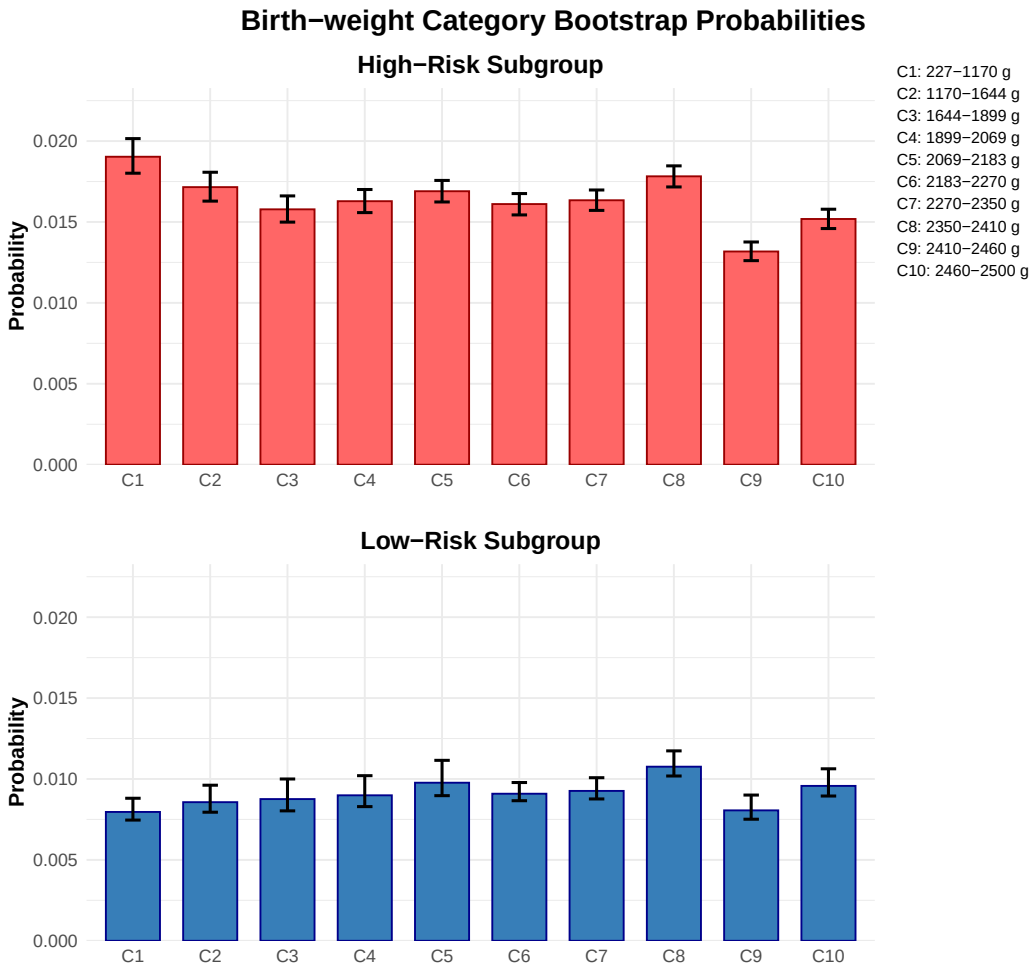


Figure 3.9: Full model (excluding the NBW column): Probability estimates and intervals are as in Fig. 3.8.

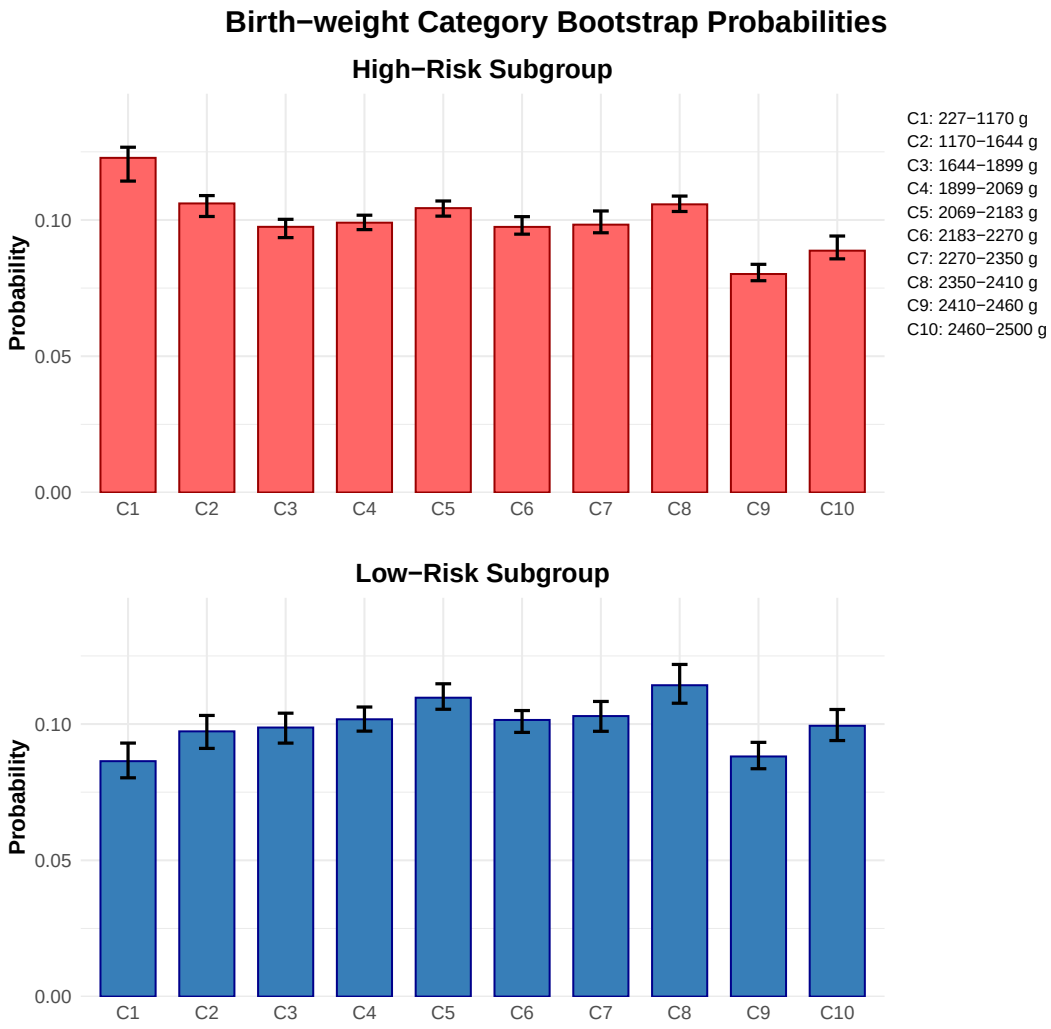


Figure 3.10: LBW-Only model (10 categories): Mean bootstrap probability for each birth-weight category in the two risk groups with 95% percentile intervals.

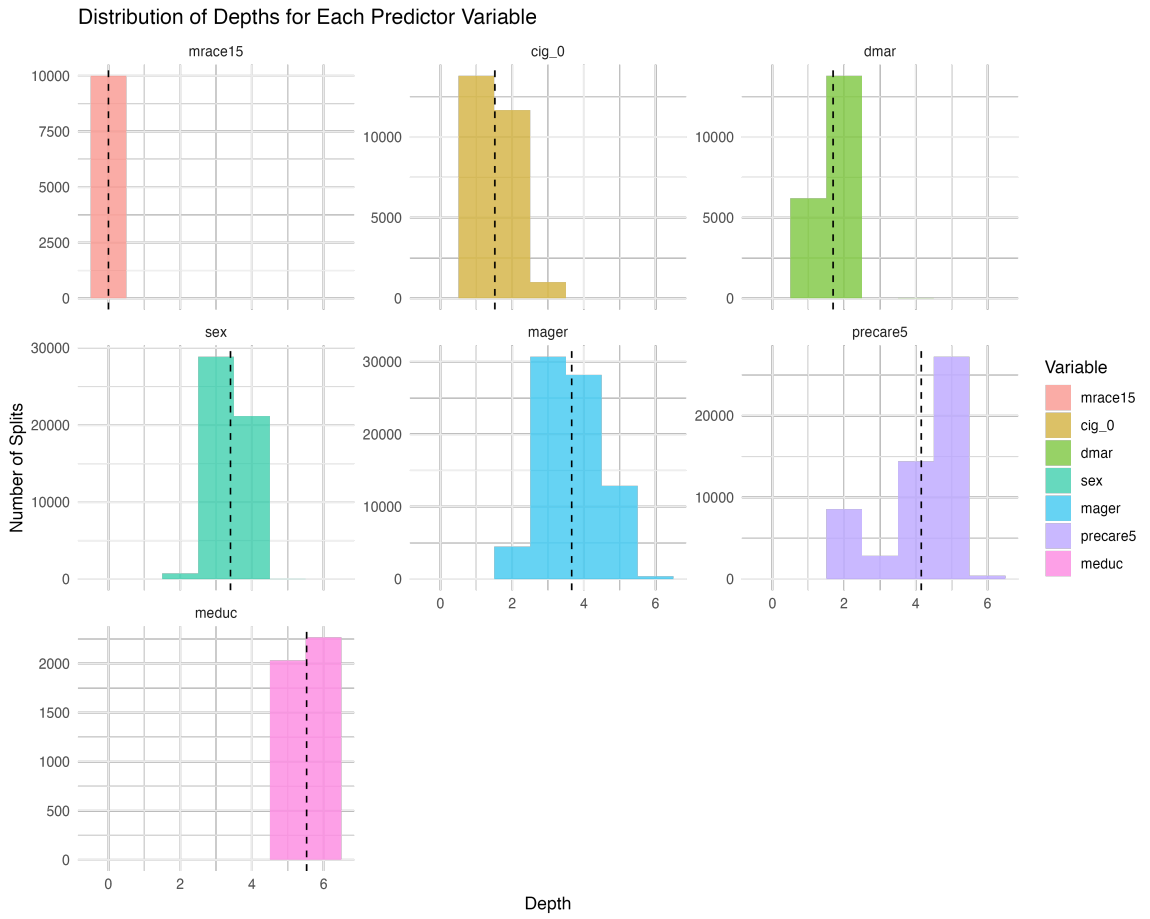


Figure 3.11: Full model: Distribution of variable depths across the ensemble. Each panel shows a histogram indicating how frequently a given variable appears at each tree depth, where depth 0 corresponds to the root node. Variables closer to the root are generally more important in the model.

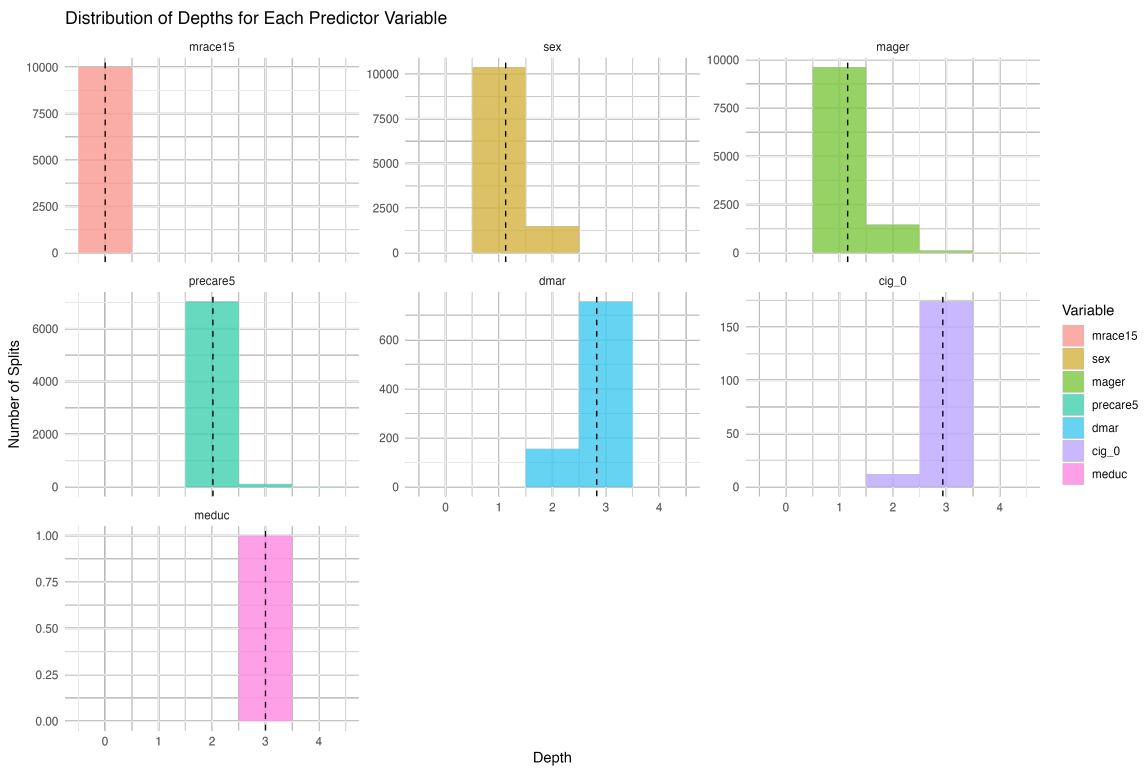


Figure 3.12: LBW-Only model: Distribution of variable depths across the ensemble. Each panel shows a histogram indicating how frequently a given variable appears at each tree depth, where depth 0 corresponds to the root node. Variables closer to the root are generally more important in the model.

Chapter 4

CONCLUSION & FUTURE WORK

This study introduces a Bayesian tree-based methodology to investigating determinants of LBW using a nationally representative dataset. By the integration of the DM likelihood into the CART framework, the model addresses both data scarcity in rare outcome classes, with addressing the need for a more flexible and interpretable modeling structure. Using historic data, the quantiles inform the priors binning procedure, creating the necessary birth-weight categories. The quantile-based categories enable the the informed prior to represent subtle gradients in the LBW-region, and enabling detection of distributional shifts given a set of predictors. Additionally, the proposed bootstrap methodology handles the consolidated counts data, while the results yield stable and reliable estimates across the ensemble.

A consistent theme in this project is that maternal race, marital status, and smoking status were dominant indicators of LBW risk. Furthermore, the restricted LBW-only model shifted the focus from socioeconomic and demographic predictors to biological and behavior-based variables. Such variables include maternal age, infant gender, and prenatal care. Note that these findings align with epidemiological literature, demonstrating the utility of our proposed modeling framework to extract and interpret clinically relevant rules from high-dimensional data.

However, the present analysis is bounded by a constrained set of binary predictors and reductionist encoding of sociodemographic and behavioral traits. In further analysis, clinically relevant information could be utilized instead of discarded by encoding predictors into binary. Further, the race predictor currently represents a coarse proxy of demographic and socioeconomic conditions in further studies should refine this important predictor to

capture more social and cultural dimensions. The most natural next step is enhancing the data with a richer predictor set of maternal-infant health and environmental indicators. Promising health related variables include: history of hypertension (Ardissino *et al.*, 2022), diabetes (Mi *et al.*, 2017), BMI (Gul *et al.*, 2020) , mental health (Nomura *et al.*, 2007), and prior pregnancy complications (Cutland *et al.*, 2017). Including these variables could enhance the model's ability to further classify at-risk subgroups, since they are all known to affect fetal growth. Incorporating environmental and contextual variables can expand the model's scope and abilities. Structural determinant such as air quality metrics, neighborhood crime rates, housing conditions, food accessibility, and proximity to prenatal services may interact with biological and behavioral risks in meaningful ways. Their inclusion would support a more holistic understanding of LBW outcomes. Also, the temporal trends of any of these variables is worth while to investigate due to the consistent annual reporting of the natality dataset.

Moreover, this modeling framework has the potential to be enhanced as a practical tool for clinical triage or public health screening. Future work should focus on adapting the modeling framework into a practitioner-friendly risk calculator suitable for intake assessments or integration into electronic health records. This would significantly enhance accessibility for health practitioners and support early identification of at-risk pregnancies.

This work contributes a flexible and interpretable modeling approach for modeling LBW determinants and lays the foundation for future interdisciplinary research that intersects statistical modeling, clinical practice, and public health policy. Expanding the set of predictors and translating the model into operational tools would be critical steps toward leveraging these insights into actionable health interventions.

URL https://medlineplus.gov/ency/article/003402.htm#:~:text=Apgar%20is%20a%20quick%20test,doing%20outside%20the%20mother's%20womb, accessed: 2025-04-26 (n.d.).

Ardissino, M., E. A. W. Slob, O. Millar, R. K. Reddy, L. Lazzari, K. H. K. Patel, D. Ryan, M. R. Johnson, D. Gill and F. S. Ng, "Maternal hypertension increases risk of preeclampsia and low fetal birthweight: Genetic evidence from a mendelian randomization study", Hypertension **79**, 3, 588–598, URL https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.121.18617 (2022).

Breiman, L., J. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees* (Chapman and Hall/CRC, New York, 1984), 1 edn., URL https://doi.org/10.1201/9781315139470.

Centre for Speech Technology Research, "The cart building algorithm (greedy search)", URL https://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/c16616.htm, accessed 2025-03-28 (n.d.).

Colen, C. G., A. T. Geronimus, J. Bound and S. A. James, "Maternal upward socioeconomic mobility and black–white disparities in infant birthweight", American Journal of Public Health URL https://pmc.ncbi.nlm.nih.gov/articles/PMC1751798 (2006).

Cutland, C. L., E. M. Lackritz, T. Mallett-Moore, A. Bardají, R. Chandrasekaran, C. Lahariya, M. I. Nisar, M. D. Tapia, J. Pathirana, S. Kochhar, F. M. Muñoz and Brighton Collaboration Low Birth Weight Working Group, "Low birth weight: Case definition and guidelines for data collection, analysis and presentation of maternal immunization safety data", Vaccine **35**, 48, 6492–6500, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5710991 (2017).

Cutland, C. L., Lackritz, E. M., Mallett-Moore, T., Bardají, A., Chandrasekaran, R., Lahariya, C., Nisar, M. I., Tapia, M. D., Pathirana, J., Kochhar, S., Muñoz, F. M., Brighton Collaboration Low Birth Weight Working Group URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5710991/, accessed: 2025-04-25 (2017).

Davison, A. C. and D. V. Hinkley, "Bootstrap methods and their application", URL https://statistique.cuso.ch/fileadmin/statistique/user_upload/BootShortHandout.pdf, lecture notes; accessed 2025-04-30 (2021).

Delcroix-Gomez, C., M.-H. Delcroix, A. Jamee, T. Gauthier, P. Marquet and Y. Aubard, "Fetal growth restriction, low birth weight and preterm birth: Effects of active or passive smoking evaluated by maternal expired co at delivery", Tobacco Induced Diseases **20**, 1–15 (2022).

Duke University, "Shrinkage (chapter 4 lecture notes)", URL https://www2.stat.duke.edu/~pdh10/Teaching/721/Materials/ch4shrinkage.pdf, accessed 2025-05-02 (n.d.).

Dunson, D. B., A. Herring and A. M. Siega-Riz, "Bayesian inference on changes in response densities over predictor clusters", Journal of the American Statistical Association **103**, 484, 1508–1517, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7059981/ (2008).

Finch, B. K., "Socioeconomic gradients and low birth weight: Empirical and policy considerations", Health Services Research **38**, 6 (Suppl.), 1819–1842, URL https://doi.org/10.1111/j.1475-6773.2003.00204.x (2003).

Goisis, A., H. Remes, K. Barclay, P. Martikainen and M. Myrskylä, "Advanced maternal age and the risk of low birth weight and preterm delivery: A within-family analysis using finnish population registers", American Journal of Epidemiology URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5860004, accessed 2025-03-27 (2017).

Gul, R., S. Iqbal, Z. Anwar, S. G. Ahdi, S. H. Ali and S. Pirzada, "Pre-pregnancy maternal bmi as a predictor of neonatal birth weight", PLOS ONE **15**, 10, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7592734 (2020).

Gundersen, G., "Deriving the dirichlet–multinomial distribution", URL https://gregorygundersen.com/blog/2020/12/24/dirichlet-multinomial, blog post; accessed 2025-03-03 (2020).

Hrba, M., M. Maciak, B. Peštová and M. Pešta, "Bootstrapping not independent and not identically distributed data", Mathematics **10**, 24, URL https://www.mdpi.com/2227-7390/10/24/4671 (2022).

Institute of Medicine, "The effectiveness of prenatal care", URL https://www.ncbi.nlm.nih.gov/books/NBK214461, accessed 2025-04-26 (1985).

Jain, P., *Algorithms for Bayesian Conditional Density Estimation on a Large Dataset*, Ph.D. thesis, Arizona State University, advised by P. R. Hahn, J. He, S. Zhou, M.-H. Kao, and S. Lan (2024).

Kaiser Family Foundation, "Racial disparities in maternal and infant health: Current status and efforts to address them", URL https://www.kff.org/racial-equity-and-health-policy/issue-brief/racial-disparities-in-maternal-and-infant-health-current-status-and-efforts-to-address-them, accessed 2025-03-03 (2025).

Kamperis, S., "Decision trees: Gini index vs. entropy", URL https://ekamperi.github.io/machine%20learning/2021/04/13/gini-index-vs-entropy-decision-trees.html, blog post; accessed 2025-04-29 (2021).

Kitsantas, P., M. Hollander and L. Li, "Using classification trees to assess low birth weight outcomes", Artificial Intelligence in Medicine **38**, 3, 275–289, URL https://www.sciencedirect.com/science/article/pii/S0933365706000583 (2006).

Kramer, M. S., "Determinants of low birth weight: Methodological assessment and meta-analysis", Bulletin of the World Health Organization **66**, 5, 663–737, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2491072/ (1987).

Lu, C., W. Zhang, X. Zheng, J. Sun, L. Chen and Q. Deng, "Combined effects of ambient air pollution and home environmental factors on low birth weight", Chemosphere **240**, 124836, URL https://www.sciencedirect.com/science/article/pii/S0045653519320752 (2020).

March of Dimes, "Low birthweight in the united states (2021–2023 average)", URL https://www.marchofdimes.org/peristats/data?reg=99&top=4&stop=42&lev=1&slev=1&obj=3, accessed 2025-02-23 (2024).

Martinson, M. L. and K. H. Choi, "Low birth weight and childhood health: The role of maternal education", Annals of Epidemiology **39**, 39–45.e2, URL https://www.sciencedirect.com/science/article/pii/S1047279718310950 (2019).

Mi, D., H. Fang, Y. Zhao and L. Zhong, "Birth weight and type 2 diabetes: A meta-analysis", Experimental and Therapeutic Medicine **14**, 6, 5313–5320 (2017).

Mimno, D., "Polya distribution exercise", URL https://mimno.infosci.cornell.edu/info6150/exercises/polya.pdf, class notes; accessed 2025-03-03 (2025).

Minka, T. P., "Estimating a dirichlet distribution", Tech. rep., Microsoft Research, URL https://tminka.github.io/papers/dirichlet/, revised 2003, 2009, 2012 (2000).

National Bureau of Economic Research, "Vital statistics natality birth data", URL https://www.nber.org/research/data/vital-statistics-natality-birth-data, data set; accessed 2025-02-23 (2024).

Nomura, Y., P. J. Wickramaratne, D. J. Pilowsky, J. H. Newcorn, B. Bruder-Costello, C. Davey, W. P. Fifer, J. Brooks-Gunn and M. M. Weissman, "Low birth weight and risk of affective disorders and selected medical illness in offspring at high and low risk for depression", Comprehensive Psychiatry **48**, 5, 470–478, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2085442 (2007).

Penn State University, "Distribution-free confidence intervals for percentiles", URL https://online.stat.psu.edu/stat415/book/export/html/835, course notes; accessed 2025-05-14 (n.d.).

Rathjens, J., A. Kolbe, J. Hölzer and C. Czado, "Bivariate analysis of birth weight and gestational age by bayesian distributional regression with copulas", Statistical Biosciences **16**, 2, 290–317, URL https://link.springer.com/article/10.1007/s12561-023-09396-4 (2024).

Stanford Medicine, "Low birth weight", URL https://www.stanfordchildrens.org/en/topic/default?id=low-birth-weight-90-P02382, accessed 2025-02-23 (2025).

stats.stackexchange.com, "Bootstrap resampling for contingency table", URL https://stats.stackexchange.com/questions/303939/bootstrap-resampling-for-contingency-table, accessed 2025-04-30 (2025a).

stats.stackexchange.com, "Gini decrease and gini impurity of children nodes", URL https://stats.stackexchange.com/questions/95839/gini-decrease-and-gini-impurity-of-children-nodes, accessed 2025-04-29 (2025b).

Therneau, T., B. Atkinson and B. Ripley, "rpart: Recursive partitioning and regression trees", URL https://cran.r-project.org/web/packages/rpart/rpart.pdf, package manual; accessed 2025-03-28 (2025).

Therneau, T. M. and E. J. Atkinson, "An introduction to recursive partitioning using the rpart routines", URL https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf, technical report; accessed 2025-03-28 (2023).

U.S. Census Bureau, "Quickfacts: United states", URL https://www.census.gov/quickfacts/fact/table/US/PST045224, accessed 2025-04-26 (2024).

Van Vliet, G., S. Liu and M. S. Kramer, "Decreasing sex difference in birth weight", Epidemiology p. 622, URL https://journals.lww.com/epidem/fulltext/2009/07000/decreasing_sex_difference_in_birth_weight.24.aspx (2009).

Wikipedia contributors, "Birth weight — Wikipedia, the free encyclopedia", URL https://en.wikipedia.org/wiki/Birth_weight, accessed 2025-04-26 (2025a).

Wikipedia contributors, "Contingency table — Wikipedia, the free encyclopedia", URL https://en.wikipedia.org/wiki/Contingency_table, accessed 2025-04-30 (2025b).

Wikipedia contributors, "Dirichlet-multinomial distribution — Wikipedia, the free encyclopedia", URL https://en.wikipedia.org/wiki/Dirichlet-multinomial_distribution, [Online; accessed 3-March-2025] (2025c).

Wikipedia contributors, "Shrinkage (statistics) — Wikipedia, the free encyclopedia", URL https://en.wikipedia.org/wiki/Shrinkage_(statistics), accessed 2025-05-01 (2025d).

Wikipedia contributors, "Summary statistics — Wikipedia, the free encyclopedia", URL https://en.wikipedia.org/wiki/Summary_statistics, accessed 2025-04-30 (2025e).