

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

ADAM KURTH

FOR THE DEGREE OF MASTERS OF SCIENCE (STATISTICS)
SCHOOL OF MATHEMATICAL & STATISTICAL SCIENCES
ARIZONA STATE UNIVERSITY

2025

ACKNOWLEDGEMENTS

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 family 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. Their 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.

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 11 birth weight categories, comprised of 10% LBW quantile categories plus one aggregated normal weight category for added LBW granularity. Classification and Regression Trees (CART) are grown 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 sex and maternal age supersede smoking and marital status as key discriminators, revealing finer biological gradients of risk. Ensemble predictions are well calibrated, and 95% bootstrap confidence intervals achieve nominal coverage.

The resulting framework combines the interpretability of decision trees with Bayesian uncertainty quantification, delivering actionable, clinically relevant insights for targeting maternal-health interventions among the most vulnerable subpopulations.

Contents

1	Introduction	1
1	Background & Problem	1
2	Literature Review & Methodology	4
1	Previous Work on Birth Weight Modeling	4
2	Current Approach & Contributions	6
3	Overview of the Birth Weight Dataset	7
4	Data Preprocessing & Feature Engineering	7
5	Marginal Dirichlet-Multinomial (DM) Likelihood	10
3	Tree-based Nonparametric Birth Weight Modeling	17
1	Introduction	17
2	Constructing the Informed Prior	18
3	DM-CART	18
4	Bootstrap Analysis & Methodology	24
4	Conclusion & Future Work	40

References	42
A Additional Material	48
1 Additional Material for Chapter 2	48
2 Additional Material for Chapter 3	48

List of Tables

2.1	Binary predictor definitions used in this study	8
2.2	Birth-weight quantile cut points and Dirichlet priors	9
3.1	Comparison of Variable Importance in Full Model and LBW-Only Model	30
3.2	Mean probability estimates $\hat{\pi}_{i,k}$ (with 95% bootstrap percentile intervals) for high- and low-risk birth-weight subgroups under the <i>full</i> model.	31
3.3	Mean probability estimates $\hat{\pi}_{i,k}$ (with 95% bootstrap percentile intervals) for high- and low-risk birth-weight subgroups under the <i>LBW-only</i> model.	31

List of Figures

3.1	Comparison of Informed Dirichlet Priors based on 2020 quantiles	19
3.2	DM Tree Structure, with normal birth weights.	20
3.3	DM Tree Structure, LBW-only model	22
3.4	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.	32
3.5	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.	33
3.6	Full model model tree structures showing splits 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.	34
3.7	LBW-only model tree structures showing splits 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.	35
3.8	Full model: Aggregated mean probability estimates by birth weight category across B bootstraps, showing the distribution of predicted probabilities for high and low risk subgroups with 95% confidence intervals.	36

3.9	LBW-only model: Aggregated mean probability estimates by birth weight category across B bootstraps, showing the distribution of predicted probabilities for high and low risk subgroups with 95% confidence intervals.	37
3.10	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.	38
3.11	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.	39

Chapter 1

Introduction

1 Background & Problem

Low birth weight (LBW), defined as a birth weight less than 2.5 kilograms [1, 2], 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 [3]. 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 [2] (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 [2]. The highly-interrelated nature of these risk factors lead to confounding, yielding an impediment for modeling birth weights 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 signify a *truly* high-risk

pregnancy for LBW.

Subsequent work in epidemiology repeatedly show that these risk factors are not found in isolation. In 2006, Kitsantas, Hollander, and Li [4] use Classification and Regression Trees (CART) developed by Breiman et al. [5] 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 maternal weight-gain during pregnancy, had sharply elevated LBW risk. This study 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 information about how far below the 2.5 kg threshold a birth weight lies. 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 thus masks important differences in the distribution of birth weights.

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 [6]. Other approaches use copula-based or density regression techniques to jointly model birth weight with related outcomes such as gestational age [7]. 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 [8]. Lower educational attainment is associated with limited healthcare access [3, 9], and environmental exposures, including tobacco use, substance abuse, and air pollution, further elevate risk by interfering with fetal growth and development [10, 11]. Inadequate prenatal care is another important factor of LBW outcomes [12]. Crucially, LBW incidence also varies sharply by racial and economic context. In the United States, the incidence rate for Black infants is double that of white newborns, comparing 14.7% to 7.1% [13]. These patterns underscore the multifactorial nature of LBW 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 distribution, while advanced Bayesian density models capture distributional details but lack intuitive clarity. This work aims to bridge the gap by developing a 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

1 Previous Work on Birth Weight Modeling

As previously mentioned, Kramer [2]’s 1987 meta-analysis identified 43 LBW determinants then categorized them into genetic, nutritional, psychosocial, etc. and assessed their effects on birth weight and prematurity. 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 [2]. While for preterm births, smoking status and low pre-pregnancy weight are strong indicators [2]. Kramer [2] concluded by stating that many potential contributors remain understudied, naming maternal work, prenatal care, and previous infections as examples. This comprehensive 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, Hollander, and Li [4] (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 [4]. The CART model revealed high-risk profiles for White and Hispanic mothers with low pregnancy weight gain, and parity and marital status defined high-risk stratification among non-smokers [4]. For instance, smoking mothers gaining less than 20 lbs were at significantly higher risk than larger

weight gain, and Black mothers formed high-risk subpopulation in some regions regardless of other factors [4]. However, predictive accuracy was only marginally better than logistic regression [4], 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, Herring, and Siega-Riz [6] (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 [6]. 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 [6]. Dunson et. al. demonstrate 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. [7] 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 normal for birth weight outcomes, and Dagum for skewed gestational age, and the copula linked them as functions of the covariates [7]. The results of this study show non-linear effects of gestational age on weight and tail-dependent associations were captured by Clayton copula [7]. The focus of bivariate outcomes here shows how distribution modeling can extend traditional regression approaches. Beyond complex copula models, Bayesian methods enrich perinatal risk modeling.

In 2024, Jain [9] 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 challenge involved.

2 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. [5] (1984) and implemented in R by Therneau et al. [14], 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 [15], 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 responses [14]. This process is applied recursively to each child node, growing a larger tree until the tree's max depth is reached or no further improvement is found [14].

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 [14]. The tree is then "trimmed" back to the best cross-validation performance [14], 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, Hollander, and Li [4]. Additionally, in line with Dunson, Herring, and Siega-Riz [6] and Jain [9], this tree-based method imposes no strict distributional assumptions on the birth weight responses allowing for non-linear 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 [16]. Additionally, the Gini index is known to suffer with data sparsity [17], and com-

pared to normal birth weight (NBW) observations, we expect a large discrepancy between the total number of observed LBW and NBW counts.

Birth weight counts observations can be safely assumed to follow a multinomial distribution, and the Dirichlet prior smooths categories not observed in the sample. 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 5 as a favorable splitting criterion for our application.

3 Overview of the Birth Weight Dataset

The primary dataset for this analysis is the 2021 Vitality Statistics Natality Birth Data [18]. 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 [9, 18]. 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 [18].

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, and geographic indicators include state, county, and metropolitan status [18]. Note that Apgar examinations examine newborn vitals five minutes after birth, observing how the newborns are handling being outside the mother’s womb [19].

4 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.

4.1 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 [9], 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 [9] and March of Dimes [13] as the primary dichotomy, *though this choice is entirely arbitrary*. According to 2024 U.S. Census Bureau [20] estimates, 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	Nativity field	Value = 1	Value = 0
Boy	<code>sex</code>	Infant is male (“M”)	Infant is female (“F”)
Married	<code>dmar</code>	Mother is married	Mother not married
Black	<code>mrace15</code>	Black / African American	Any other race
Over33	<code>mager</code>	Maternal age > 33 yr	Maternal age \leq 33 yr
HighSchool	<code>medu</code>	High-school education completed	Otherwise
FullPrenatal	<code>prenatal</code>	Adequate prenatal care	Inadequate / none
Smoker	<code>cig_0</code>	Any prenatal smoking	No smoking

4.2 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 rows. 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.

4.3 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 a given number of birth weight counts of a given class and quantile. The specific cut-point values and their prior assignment are shown in the following table:

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

Quantile	Type 1: LBW + Normal		Type 2: LBW only	
	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		

As seen in Table 2.2 the two types include: low and normal birth weights (NBW), and the other restricted to LBW, where NBW is defined as any birth weight observation greater than 2.5 kg [21]. 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 eleventh 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 construction is heavily skewed for the low and NBW with probability of 91.67%, while the second 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 in the dataset. This approach prevents scarcity in any one category, and the data will be called *counts data* from here forward.

5 Marginal Dirichlet-Multinomial (DM) Likelihood

5.1 Introduction

In the previous section, we established the discretized LBW quantile categories for this study. Resulting in 10 LBW categories of quantile increments of 10% and the aggregated eleventh category for NBW. Modeling the distribution of LBW count outcomes must require handling categorical partitions of the LBW 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 that are not observed in the sample. If we disregard this issue, 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) they appear in the data. By introducing the Dirichlet prior, smoothing adds extra variability in the likelihood, 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 [22, 23, 24]. The Dirichlet's overdispersion, effectively injects "pseudo-counts" or zero-inflating prior observations in all birth weight categories [22, 25]. This ensured that the marginal likelihood, or evidence, for any category remains strictly positive, i.e. non-zero probability assignment for unobserved categories [25]. Here, $\alpha = (\alpha_1, \dots, \alpha_K)$ 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 is

discussed in Section 5.3.

The Dirichlet prior’s initial uncertainty and the actualized count observations following a multinomial distribution, the choice of the DM likelihood is an appropriate choice for birth weight modeling. The Bayesian approach 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.

5.2 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) and 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 $\mathbf{Y} := [n_{i,k}]_{i=1,\dots,N}^{k=1,\dots,K}$ of dimensions $N \times K$, where $n_{i,k}$ is a number of birth observations for class i , and quantile k . For class $i = 1, \dots, N$, each \mathbf{x}_i of \mathbf{X} is a 7-dimensional feature vector representing a unique combination of predictors. The corresponding row $\mathbf{y}_i = (n_{i,1}, \dots, n_{i,K})$ in \mathbf{Y} is of length K of counts observations for $k = 1, \dots, K$ birth weight categories. 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 4, which dramatically consolidate the dataset into the counts data format. 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 $N_i = 6 + 3 + 1 = 10$. By using the count vectors for each predictor vector, we can apply the DM likelihood model for multivariate counts data.

5.3 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 $\mathbf{y}_i = (n_{i,1}, \dots, n_{i,K})$ with the total number of counts in class i represented as $N_i = \sum_{k=1}^K n_{i,k}$. In the classic multinomial

probability mass function, the factor

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

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 to counts data, we no longer model the possible orderings of N_i births, only the counts vector \mathbf{y}_i . Because this coefficient is a constant with respect to the category probability vector θ_i [25], it plays no role in the likelihood-ratio comparisons and is therefore 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:

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

5.4 Derivation

The hierarchical model structure is as follows:

$$\begin{aligned} \mathbf{x}_i &= \text{dummy-encoded predictor vector for class } i, \quad i = 1, \dots, 128 \\ \mathbf{y}_i \mid \theta_i, \mathbf{x}_i &\sim \text{AdjustedMultinomial}(N_i, \theta_i) \\ \theta_i &\sim \text{Dirichlet}(\alpha = (\alpha_1, \dots, \alpha_K)) \quad (\text{prior}) \\ \theta_i \mid \mathbf{y}_i, \mathbf{x}_i &\sim \text{Dirichlet}(\alpha + \mathbf{y}_i) \quad (\text{posterior}) \end{aligned}$$

Where $N_i = \sum_{k=1}^K n_{i,k}$ is the row total and $\theta_i = (\theta_{i,1}, \dots, \theta_{i,K})$ 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 $(n_{i,1}, \dots, n_{i,K})$ into (n_1, \dots, n_K) , where n_1 is naturally interpreted as the first 10% quantile. Likewise let $\theta = (\theta_1, \dots, \theta_K)$ 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:

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

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 [25]. Given θ , the probability of observing a specific count outcome follows the adjusted multinomial likelihood in Section 5.3. This is the likelihood of the category k for a given θ_k .

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

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:

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

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 [22]. 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 $(\alpha_1 + n_1, \alpha_2 + n_2, \dots, \alpha_K + n_K)$ 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 [25]:

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

Taking the log transform yields the equivalent form of which we directly implement Equation 2.3 in our code:

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

The equations above can be recognized as the DM distribution, sometimes called Dirichlet Compound Multinomial or Pólya's urn distribution [22]. 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 all categories sum to 1, yielding the normalization factor across joint observation counts [23, 25]. 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 that guarantees $p(\mathbf{y} \mid \boldsymbol{\alpha}) > 0$ for all possible outcomes [22].

5.5 Splitting with Adjusted DM Likelihood in CART

Equation 2.3 is the final log marginal DM likelihood (evidence) that will be implemented in CART. Now we can understand how CART uses this criterion for evaluating splits within 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 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 $\mathcal{L}_{\text{node}}$ 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: $\mathcal{L}_{\text{parent}} = \log p(\mathbf{y}_{\text{parent}} \mid \boldsymbol{\alpha}, \mathbf{x})$ then, $\mathcal{L}_{\text{left}} = \log p(\mathbf{y}_{\text{left}} \mid \boldsymbol{\alpha}, \mathbf{x})$ and $\mathcal{L}_{\text{right}} = \log p(\mathbf{y}_{\text{right}} \mid \boldsymbol{\alpha}, \mathbf{x})$ and calculate the improvement gain of the split as:

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

Essentially, if $\Delta \mathcal{L}$ is positive the split has results in a improvement where the DM criterion favors the split with more homogeneity. If the multinomial coefficient was included then for any partition, \mathcal{L} would directly reward partitions with

the larger number of possible orderings of the data. As noted in Section 5.3, 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 $\Delta\mathcal{L}$, 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 and thus no split is chosen. Conversely, if there's a predictor where one node gets all LBW observations, and the other all normal birth weights, the combined likelihood of child nodes would be much higher than the parent. This would cause a large $\Delta\mathcal{L}$ 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

1 Introduction

We model birth weight risk with a tree-based, nonparametric approach focusing on the LBW region while still modeling the full spectrum of birth weights. 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 affect 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 Dirichlet hyperparameters from the *previous* year's data (2020) and apply them to the unchanged 2021 dataset. Figure 3.1 shows the resulting prior vector α . The full model (left) shows expected proportion of observations in the NBW category, and the LBW-only (right) shows a 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 sex (`sex`)—

are given to CART, searching recursively for the largest improvement gain. To confirm stability and reliability of the split predictors we will provide a 10,000 bootstrap ensemble and compare the variable selection across the trees. This will be further elaborated on in Sections ??

2 Constructing the Informed Prior

2.1 Quantile Cut-points & Informed Prior from 2020 Data

To avoid "double dipping", or using the same observations for both the model fit and setting its prior, we derive the LBW quantile cut-points and Dirichlet hyperparameters from the *previous* year's dataset (2020) and apply them to the unchanged 2021 dataset. All birth weights less than the threshold of 2.5 kg are divided into 10 equal-frequency categories, or deciles, comprising the 10% quantiles mentioned earlier. The LBW 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. A comparison of LBW-only model omits this last category so that its attention centers around variation *within* the LBW region in Section 3.3.

From the 2020 proportions we construct the Dirichlet prior, $\alpha = (\alpha_1, \dots, \alpha_K)$. Figure 3.1 depicts the resulting priors: with the left panel highlighting the strength toward NBW in the full model, whereas the right panel shows the uniform prior among LBW categories in the LBW-only model. The national birth weight distributions vary little year-over-year, the 2020 data serves as an appropriate proxy for 2021.

3 DM-CART

3.1 Custom Objective Function & Splitting

We fitted CART using `rpart`, in R [26], replacing the default Gini index with the adjusted marginal log DM likelihood or simply the log likelihood for brevity. Serving as our objective function, its output scores based on the reduction in de-

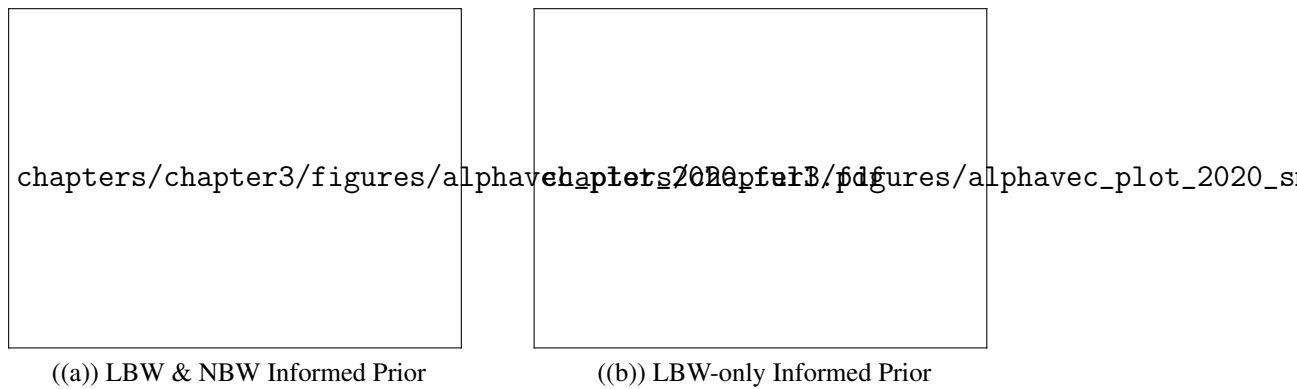



Figure 3.1: Comparison of Informed Dirichlet Priors based on 2020 quantiles

viance of a given split, i.e. negative log likelihood output [26, 14]. When splitting, `rpart` computes the left and right split deviance calculation as in Section 5.5. For instance, suppose we propose a split on smoking status. The objective function evaluates separating the data into smoking and non-smoking mothers is "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.

3.2 Tree Results and Insights: Full Model

For the full model, Figure 3.2 shows the hierarchical structure of the splits found, and Figure 3.4 ranks improvement gain for each split. The full model splits as follows:



chapters/chapter3/figures/dm_tree2021.pdf

Figure 3.2: DM Tree Structure, with normal birth weights.

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 ($\text{mrace15}=1$) from all other mothers ($\text{mrace15}=0$). This split yields the largest difference in birth weight profile and is shown to be the most informative predictor. This split aligns with documented discrepancies of race, playing a critical role in LBW incidence [27, 28]. 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 subpopulations are among unmarried Black smokers where $\text{mrace15} = 1, \text{cig_0} = 1, \text{dmar} = 0$. 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 demographic splits [27, 28, 29] and Figure 3.2 and Figure 3.4 visualize these results. Further splitting down the branch of Black smokers provides further nuance of the highest-risk subgroups ($\text{mrace15} = 1, \text{cig_0} = 1, \text{dmar} = 0$). This node splits on infant gender,

where on average, female infants weight less than male infants [30].

Likewise, the lowest-risk subpopulations are where $\text{mrace15}=0$ — recall that smoking status is considered after marital status of the mother. Despite the ordering of ranked, smoking is a direct determinant of LBW incidence [29]. After socioeconomic and demographic variables are split on, the model emphasizes more biological and genetic predictors, namely gender, prenatal care, and maternal age. The the lowest-risk groups where $\text{mrace15} = 0, \text{dmar} = 1, \text{cig}_0 = 0$, the final predictor that is considered is age. Moreover, this branch has a depth of 4 while the highest-risk subgroups have a depth of 6 or 7.

Surprisingly, the only case where education status (meduc) of the mother was used was if the mother was Black, non-smoker, below 33 years old, without adequate prenatal care and female newborn, (i.e. $\text{mrace15} = 1, \text{cig}_0 = 0, \text{dmar} = 0, \text{sex} = 0, \text{mager} = 0, \text{precare5} = 0$). Given that education has been noted by many [31] to have an effect on socioeconomic conditions, namely earnings, which might not be as strong of a predictor as initially thought.

The full model tree 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 3.4. 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.

3.3 Tree Results and Insights: LBW-Only

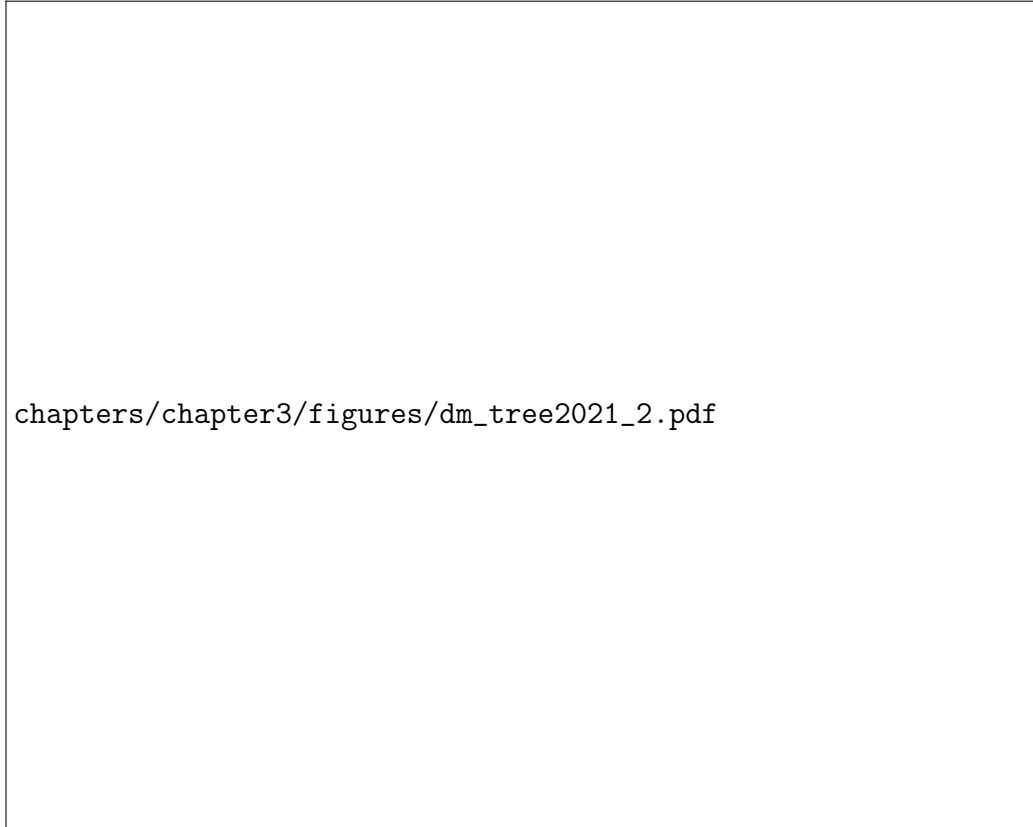


Figure 3.3: DM Tree Structure, LBW-only model

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 the restriction that the LBW-only model brings. 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 in the Black branch and maternal age in the non-Black 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 3.4 and 3.5 order predictors by the sum of total deviance each eliminates across all splits, showing the contrast between predictor usage. The barplots reflect 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 association with response [15].

3.4 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 3.6 and 3.7 visualize the trees.

As the depths increase for the full model (Figure 3.6), 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 birth weight outcomes.

The LBW-restricted trees (Figure 3.7) show a similar growth in complexity from 4 terminal nodes at depth 2, to 15 at depth of 5, and the number of predictors likewise rises from 3 to 6. The number of terminal nodes reflects the more homogeneity of the LBW-only restricted data, consequently considering less splits and returning less leaves compared to the full model. The top splitter among the LBW-only tree was shown to be *marital status* rather than race. The subsequent splits consider race, infant gender, maternal age, and prenatal care. Notably, at depth 4 race was shown to be of significantly less importance compared to the full model. For unmarried mothers, race is considered at the second split on the left branch while on the right branch, for *married* mothers it is considered last and is the terminal node behind infant gender, maternal age, and prenatal care. These results suggest among LBW newborns, marital status plays a critical role in child-development [32]. Crucially, *cig_0* is never selected, indicating that smoking does not differentiate outcomes once infants are already below 2.5 kg.

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 Sec-

tion 3.3. Due to this difference, we will employ a two-tier bootstrap procedure to confirm stability of variable selection and importance.

4 Bootstrap Analysis & Methodology

4.1 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. Each replicated resampling perturbs the data into two levels that mirror 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.

4.2 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 breach the key independence assumption that underpins case-resampling [33, slide 47] [34]. 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 [35]. De-aggregating these frequencies into individual observations before performing bootstrap resampling is required to preserve the within-row dependence, preserving the data structure [36], precisely because each row is a vector of summary statistics [37]. Further, when sampling any row, the counts vector has fixed proportions of LBW and NBW counts, thereby provid-

ing no within-row randomness. The NBW proportion greatly exceeds that of the LBW counts, overshadowing 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 2.

4.3 Methodology & Procedure

Formally, the two-tier bootstrap resampling procedure is defined here. First, T defines the grand total counts. $\mathbf{p} = (p_1, \dots, p_N)$ are the empirical proportions across N classes. The *bootstrap probability estimates* for class i across K categories are represented as, $\hat{\pi}_i = (\hat{\pi}_{i,1}, \dots, \hat{\pi}_{i,K})$, these estimates follow the closed-form Equation 2.1. Additionally, the predictor matrix \mathbf{X} remains fixed; only response counts are resampled.

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

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

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

whose mean estimate,

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

represent the bootstrap probability estimates for i, k . Importantly, these K estimates are fixed. The posterior distribution infers about $\theta_{i,k}$ under Equation 2.1, whereas the shrinkage estimate $\hat{\pi}_{i,k}$ is referred to as fixed under a multinomial distribution in Tier 4.3 [38, 39].

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.

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

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

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

Since $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i,1}, \dots, \hat{\pi}_{i,K})$ is plugged-in and held fixed here, each vector $\hat{\boldsymbol{\pi}}_i$ under class i is referred to as the bootstrap probability estimates for K birth weight categories. Each bootstrap replicate inherits prior information via $\boldsymbol{\alpha}$, 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 3.10, 3.11. 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 posterior probabilities of variable selection under the DM hierarchy.

Note that for each class i , we compute the posterior mean probability estimates $\hat{\boldsymbol{\pi}}_i$ and hold this vector as fixed when we generate the replicate counts $\tilde{\mathbf{y}}_i \sim \text{Multinomial}(n_i^*, \hat{\boldsymbol{\pi}}_i)$. Rather than resampling a new $\boldsymbol{\theta}_i^{(b)} \sim \text{Dirichlet}(\boldsymbol{\alpha} + \mathbf{y}_i)$ 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 $\boldsymbol{\alpha}$ from being counted twice.

4.4 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 modeling 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 3.4) that racial disparities represent the most prominent signal in the data. Figures 3.10 and 3.11 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 3.10, 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 changed substantially, reflecting a shift in the model's 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 3.10, 3.11. 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.

4.5 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, its distribution-free [40]. 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 [40]. Figures 3.8 and 3.9 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 ($\text{mrace15} = 1$, $\text{dmar} = 0$, $\text{cig_0} = 1$, $\text{sex} = 0$, $\text{mager} = 0$, $\text{prenatal} = 0$, $\text{meduc} = 0$).
- *Low-risk* ($i = 28$): married, non-Black, non-smoking mothers aged 33+, \geq High-School education, adequate prenatal care, delivering male infants ($\text{mrace15} = 0$, $\text{dmar} = 1$, $\text{cig_0} = 0$, $\text{sex} = 1$, $\text{mager} = 1$, $\text{prenatal} = 1$, $\text{meduc} = 1$).

In the full model, Figure 3.8 and Table 3.2, both profiles have a very high predicted probability of delivering a NBW infant, yet the high-risk profile's NBW chance (83.6%) is roughly seven percentage points lower than the low-risk profile (90.9%). The modest absolute differences corresponds to the pronounced risks across adverse categories. In the most severe LBW category ($k = 1$) the highest probability is more than double that of the low-risk subgroup (1.9% vs. 0.8%). Percentile intervals are extremely narrow ($\approx \pm 0.002$) indicating remarkable stability across all bootstrap replicates.

Likewise, the LBW-only model in Figure 3.9 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 $k = 1$), 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 $k = 8$ 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
<i>Initial Split Variable</i>		
mrace15	1.0000	1.0000
<i>Variable Frequency</i>		
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; confidence-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; confidence-limit columns remain on the $[0, 1]$ scale. Estimates are based on B bootstrap replicates.

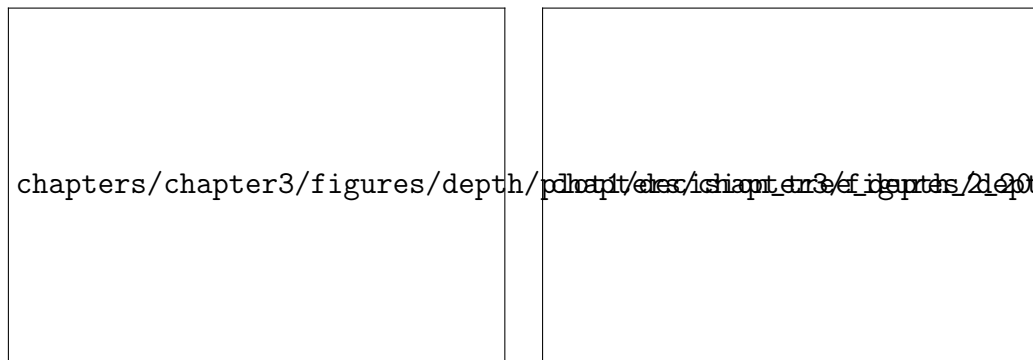


chapters/chapter3/figures/improvement/tree_split_contribution_top20_Full Model.p

Figure 3.4: 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.

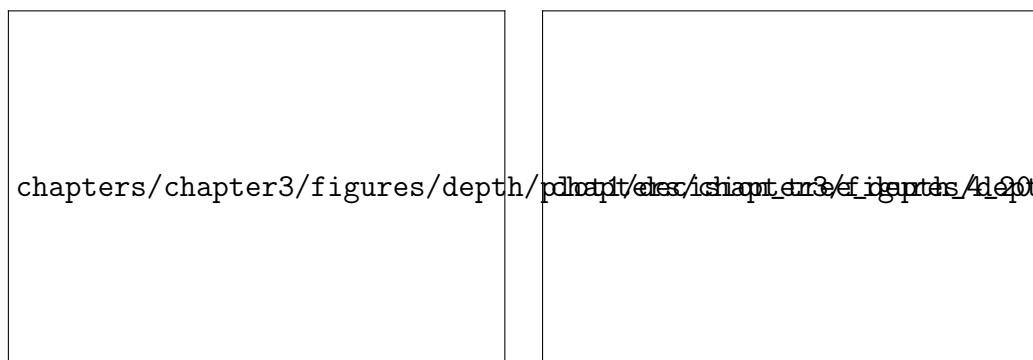


Figure 3.5: LBW-only Model Ranked Improvement. Rankings represent summed reduction in deviance (improvement in model fit) across all nodes where each variable is used for splitting in the tree. Plot displays all ranked variables.



Maximum depth = 2

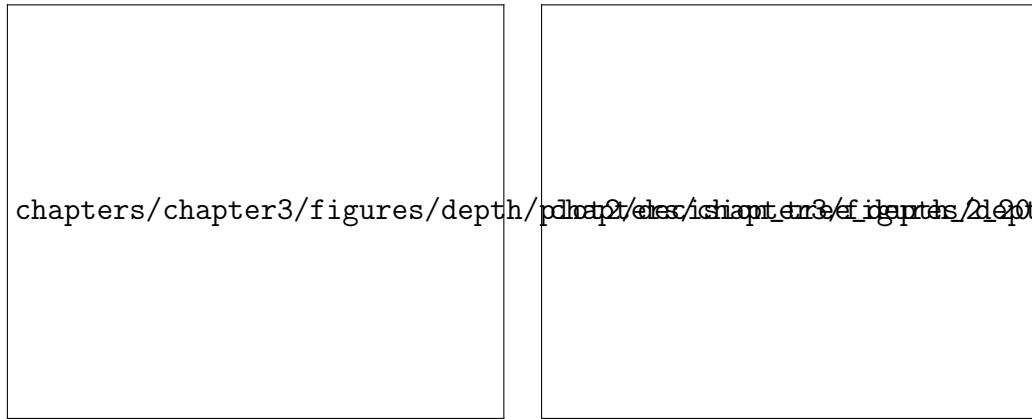
Maximum depth = 3



Maximum depth = 4

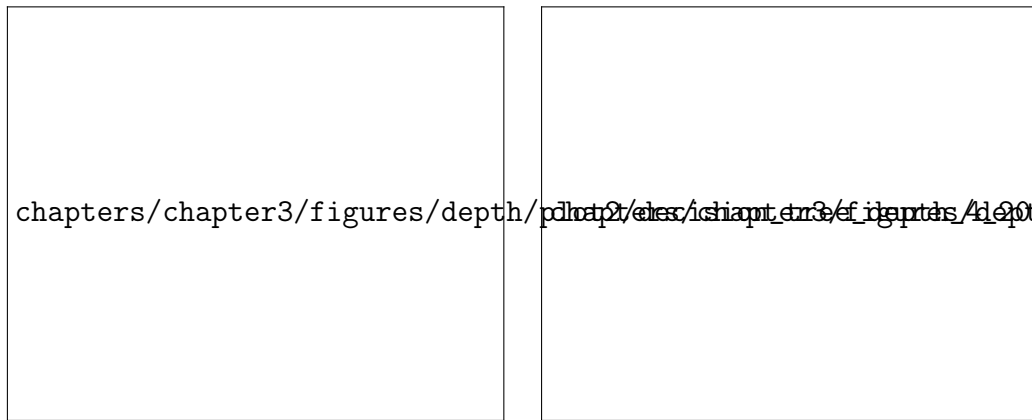
Maximum depth = 5

Figure 3.6: Full model model tree structures showing splits 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 3.7: LBW-only model tree structures showing splits 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.



Figure 3.8: Full model: Aggregated mean probability estimates by birth weight category across B bootstraps, showing the distribution of predicted probabilities for high and low risk subgroups with 95% confidence intervals.

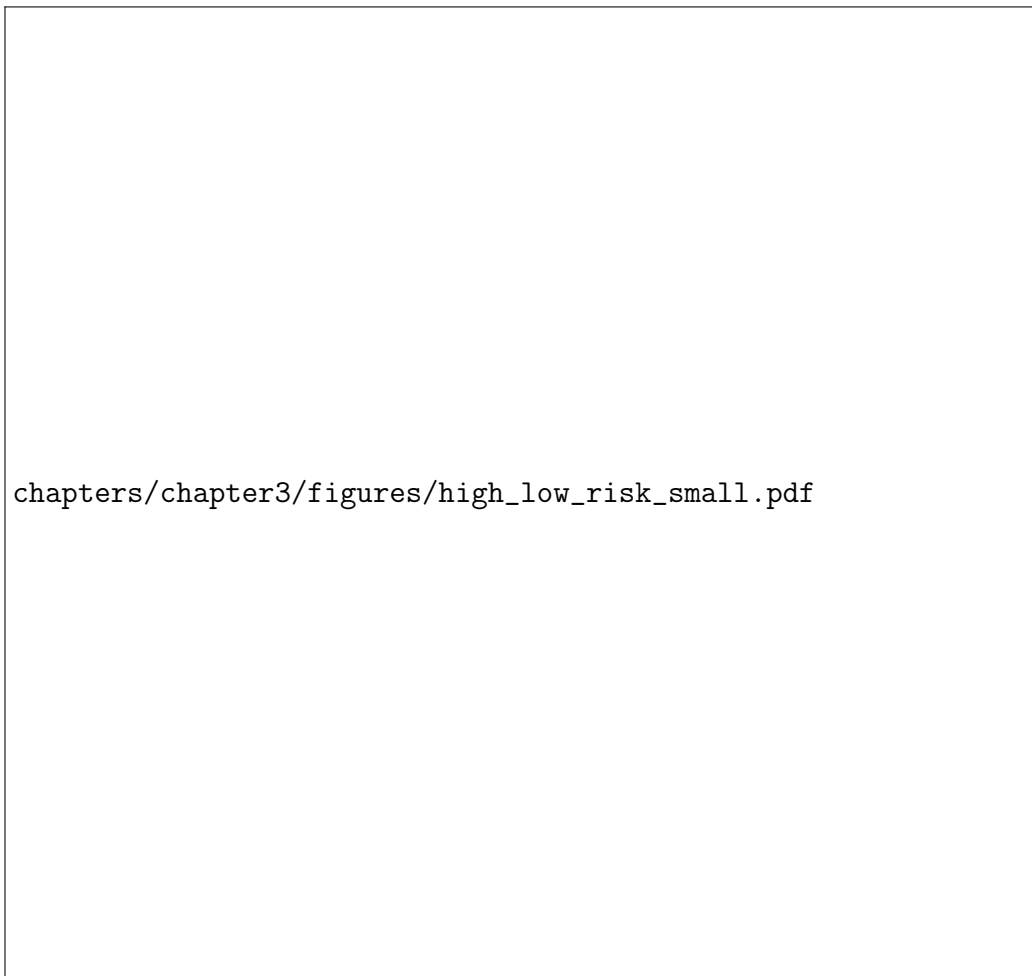


Figure 3.9: LBW-only model: Aggregated mean probability estimates by birth weight category across B bootstraps, showing the distribution of predicted probabilities for high and low risk subgroups with 95% confidence intervals.



Figure 3.10: 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 3.11: 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 and the need for a more flexible, 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 detect 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 and in further studies this should be refined 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 [41], diabetes [42], BMI [43], mental health [44], and prior pregnancy complications [45]. 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 extend 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, linking individual and population-level interventions. 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.

References

- [1] 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. In: (2017). Accessed: 2025-04-25. DOI: <https://doi.org/10.1016/j.vaccine.2017.01.049>. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5710991/>.
- [2] M. S. Kramer. “Determinants of Low Birth Weight: Methodological Assessment and Meta-analysis”. In: *Bulletin of the World Health Organization* 66.5 (1987), pp. 663–737. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC2491072/>.
- [3] B. K. Finch. “Socioeconomic Gradients and Low Birth Weight: Empirical and Policy Considerations”. In: *Health Services Research* 38.6 (Suppl.) (2003), pp. 1819–1842. URL: <https://doi.org/10.1111/j.1475-6773.2003.00204.x>.
- [4] P. Kitsantas, M. Hollander, and L. Li. “Using Classification Trees to Assess Low Birth Weight Outcomes”. In: *Artificial Intelligence in Medicine* 38.3 (2006), pp. 275–289. DOI: 10.1016/j.artmed.2006.03.008. URL: <https://www.sciencedirect.com/science/article/pii/S0933365706000583>.
- [5] Leo Breiman et al. *Classification and Regression Trees*. 1st ed. New York: Chapman and Hall/CRC, 1984. DOI: 10.1201/9781315139470. URL: <https://doi.org/10.1201/9781315139470>.
- [6] D. B. Dunson, A. Herring, and A. M. Siega-Riz. “Bayesian Inference on Changes in Response Densities over Predictor Clusters”. In: *Journal of the American Statistical Association* 103.484 (2008), pp. 1508–1517. DOI: 10.1198/016214508000001039. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7059981/>.

- [7] J. Rathjens et al. “Bivariate Analysis of Birth Weight and Gestational Age by Bayesian Distributional Regression with Copulas”. In: *Statistical Biosciences* 16.2 (2024), pp. 290–317. DOI: 10.1007/s12561-023-09396-4. URL: <https://link.springer.com/article/10.1007/s12561-023-09396-4>.
- [8] A. Goisis et al. “Advanced Maternal Age and the Risk of Low Birth Weight and Preterm Delivery: A Within-Family Analysis Using Finnish Population Registers”. In: *American Journal of Epidemiology* (2017). Accessed 2025-03-27. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5860004>.
- [9] Palak Jain. “Algorithms for Bayesian Conditional Density Estimation on a Large Dataset”. Advised by P. R. Hahn, J. He, S. Zhou, M.-H. Kao, and S. Lan. PhD thesis. Arizona State University, Aug. 2024.
- [10] Stanford Medicine. “Low Birth Weight”. In: (2025). Accessed 2025-02-23. URL: <https://www.stanfordchildrens.org/en/topic/default?id=low-birth-weight-90-P02382>.
- [11] C. Lu et al. “Combined Effects of Ambient Air Pollution and Home Environmental Factors on Low Birth Weight”. In: *Chemosphere* 240 (2020), p. 124836. DOI: 10.1016/j.chemosphere.2019.124836. URL: <https://www.sciencedirect.com/science/article/pii/S0045653519320752>.
- [12] Institute of Medicine. *The Effectiveness of Prenatal Care*. Accessed 2025-04-26. 1985. URL: <https://www.ncbi.nlm.nih.gov/books/NBK214461>.
- [13] March of Dimes. *Low Birthweight in the United States (2021–2023 Average)*. Accessed 2025-02-23. 2024. URL: <https://www.marchofdimes.org/peristats/data?reg=99&top=4&stop=42&lev=1&slev=1&obj=3>.
- [14] Terry M. Therneau and Elizabeth J. Atkinson. “An Introduction to Recursive Partitioning Using the RPART Routines”. In: (2023). Technical report; accessed 2025-03-28. URL: <https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf>.
- [15] Centre for Speech Technology Research. *The CART Building Algorithm (Greedy Search)*. Accessed 2025-03-28. n.d. URL: https://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/c16616.htm.

- [16] stats.stackexchange.com. *Gini Decrease and Gini Impurity of Children Nodes*. Accessed 2025-04-29. 2025. URL: <https://stats.stackexchange.com/questions/95839/gini-decrease-and-gini-impurity-of-children-nodes>.
- [17] Stathis Kamperis. *Decision Trees: Gini Index vs. Entropy*. Blog post; accessed 2025-04-29. 2021. URL: <https://ekamperi.github.io/machine%20learning/2021/04/13/gini-index-vs-entropy-decision-trees.html>.
- [18] National Bureau of Economic Research. *Vital Statistics Natality Birth Data*. Data set; accessed 2025-02-23. 2024. URL: <https://www.nber.org/research/data/vital-statistics-natality-birth-data>.
- [19] Accessed: 2025-04-26. n.d. URL: <https://medlineplus.gov/ency/article/003402.htm#:~:text=Apgar%20is%20a%20quick%20test,doing%20outside%20the%20mother's%20womb>.
- [20] U.S. Census Bureau. *QuickFacts: United States*. Accessed 2025-04-26. 2024. URL: <https://www.census.gov/quickfacts/fact/table/US/PST045224>.
- [21] Wikipedia contributors. *Birth Weight* — *Wikipedia, The Free Encyclopedia*. Accessed 2025-04-26. 2025. URL: https://en.wikipedia.org/wiki/Birth_weight.
- [22] David Mimno. *Polya Distribution Exercise*. Class notes; accessed 2025-03-03. 2025. URL: <https://mimno.infosci.cornell.edu/info6150/exercises/polya.pdf>.
- [23] Gregory Gundersen. *Deriving the Dirichlet–Multinomial Distribution*. Blog post; accessed 2025-03-03. 2020. URL: <https://gregorygundersen.com/blog/2020/12/24/dirichlet-multinomial>.
- [24] Thomas P. Minka. *Estimating a Dirichlet Distribution*. Tech. rep. Revised 2003, 2009, 2012. Microsoft Research, 2000. URL: <https://tminka.github.io/papers/dirichlet/>.
- [25] Wikipedia contributors. *Dirichlet-multinomial distribution* — *Wikipedia, The Free Encyclopedia*. [Online; accessed 3-March-2025]. 2025. URL: https://en.wikipedia.org/wiki/Dirichlet-multinomial_distribution.
- [26] Terry Therneau, Beth Atkinson, and Brian Ripley. “rpart: Recursive Partitioning and Regression Trees”. In: (2025). Package manual; accessed 2025-03-28. URL: <https://cran.r-project.org/web/packages/rpart/rpart.pdf>.

- [27] Kaiser Family Foundation. *Racial Disparities in Maternal and Infant Health: Current Status and Efforts to Address Them*. Issue Brief. Accessed 2025-03-03. 2025. 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>.
- [28] C. G. Colen et al. “Maternal Upward Socioeconomic Mobility and Black–White Disparities in Infant Birthweight”. In: *American Journal of Public Health* (2006). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC1751798>.
- [29] C. Delcroix-Gomez et al. “Fetal Growth Restriction, Low Birth Weight and Preterm Birth: Effects of Active or Passive Smoking Evaluated by Maternal Expired CO at Delivery”. In: *Tobacco Induced Diseases* 20 (2022), pp. 1–15. DOI: 10.18332/tid/152111.
- [30] G. Van Vliet, S. Liu, and M. S. Kramer. “Decreasing Sex Difference in Birth Weight”. In: *Epidemiology* (2009), p. 622. URL: https://journals.lww.com/epidem/fulltext/2009/07000/decreasing_sex_difference_in_birth_weight.24.aspx.
- [31] M. L. Martinson and K. H. Choi. “Low Birth Weight and Childhood Health: The Role of Maternal Education”. In: *Annals of Epidemiology* 39 (2019), 39–45.e2. DOI: 10.1016/j.annepidem.2019.09.006. URL: <https://www.sciencedirect.com/science/article/pii/S1047279718310950>.
- [32] A. Merklinger-Gruchala and M. Kapiszewska. “Marital Status, Father Acknowledgement and Birth Outcomes: Does Maternal Education Matter?” In: *International Journal of Environmental Research and Public Health* 20 (2023), p. 4868. DOI: 10.3390/ijerph20064868. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10048939>.
- [33] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Lecture notes; accessed 2025-04-30. 2021. URL: https://statistique.cuso.ch/fileadmin/statistique/user_upload/BootShortHandout.pdf.
- [34] M. Hrba et al. “Bootstrapping Not Independent and Not Identically Distributed Data”. In: *Mathematics* 10.24 (2022). DOI: 10.3390/math10244671. URL: <https://www.mdpi.com/2227-7390/10/24/4671>.

- [35] Wikipedia contributors. *Contingency Table* — *Wikipedia, The Free Encyclopedia*. Accessed 2025-04-30. 2025. URL: https://en.wikipedia.org/wiki/Contingency_table.
- [36] stats.stackexchange.com. *Bootstrap Resampling for Contingency Table*. Accessed 2025-04-30. 2025. URL: <https://stats.stackexchange.com/questions/303939/bootstrap-resampling-for-contingency-table>.
- [37] Wikipedia contributors. *Summary Statistics* — *Wikipedia, The Free Encyclopedia*. Accessed 2025-04-30. 2025. URL: https://en.wikipedia.org/wiki/Summary_statistics.
- [38] Wikipedia contributors. *Shrinkage (Statistics)* — *Wikipedia, The Free Encyclopedia*. Accessed 2025-05-01. 2025. URL: [https://en.wikipedia.org/wiki/Shrinkage_\(statistics\)](https://en.wikipedia.org/wiki/Shrinkage_(statistics)).
- [39] Duke University. *Shrinkage (Chapter 4 Lecture Notes)*. Accessed 2025-05-02. n.d. URL: <https://www2.stat.duke.edu/~pdh10/Teaching/721/Materials/ch4shrinkage.pdf>.
- [40] Penn State University. *Distribution-Free Confidence Intervals for Percentiles*. Course notes; accessed 2025-05-14. n.d. URL: <https://online.stat.psu.edu/stat415/book/export/html/835>.
- [41] M. Ardisino et al. “Maternal Hypertension Increases Risk of Preeclampsia and Low Fetal Birthweight: Genetic Evidence from a Mendelian Randomization Study”. In: *Hypertension* 79.3 (2022), pp. 588–598. DOI: 10.1161/HYPERTENSIONAHA.121.18617. URL: <https://www.ahajournals.org/doi/10.1161/HYPERTENSIONAHA.121.18617>.
- [42] D. Mi et al. “Birth Weight and Type 2 Diabetes: A Meta-analysis”. In: *Experimental and Therapeutic Medicine* 14.6 (2017), pp. 5313–5320. DOI: 10.3892/etm.2017.5234.
- [43] R. Gul et al. “Pre-pregnancy Maternal BMI as a Predictor of Neonatal Birth Weight”. In: *PLOS ONE* 15.10 (2020). DOI: 10.1371/journal.pone.0240748. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7592734>.
- [44] Y. Nomura et al. “Low Birth Weight and Risk of Affective Disorders and Selected Medical Illness in Offspring at High and Low Risk for Depression”. In: *Comprehensive Psychiatry* 48.5 (2007), pp. 470–478. DOI: 10.1016/j.comppsy.2007.04.005. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC2085442>.

- [45] C. L. Cutland et al. “Low Birth Weight: Case Definition and Guidelines for Data Collection, Analysis and Presentation of Maternal Immunization Safety Data”. In: *Vaccine* 35.48 (2017), pp. 6492–6500. DOI: 10.1016/j.vaccine.2017.01.049. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5710991>.

Appendix A

Additional Material

1 Additional Material for Chapter 2

2 Additional Material for Chapter 3