Exploring Severe Maternal Morbidity in the United States Using Supervised and Unsupervised Learning

Stephanie Diamond

DATA 71200

For my projects, I chose to examine Severe Maternal Morbidity in the United States in 2024. Severe Maternal Morbidity, or SMM, refers to unexpected, life-threatening complications that occur during pregnancy, labor, delivery, or the postpartum period (CDC, 2021). The Centers for Disease Control and Prevention (CDC) define SMM broadly, using a set of 21 clinical indicators that capture both serious diagnoses such as eclampsia or sepsis, and critical medical interventions, including blood transfusion or emergency hysterectomy (CDC, 2021).

Data Source and Preparation

To study SMM, I use the 2024 Natality Public Use File, produced by the National Center for Health Statistics (NCHS) as part of the National Vital Statistics System (NVSS). This dataset includes all registered births in the United States, and includes variables to desribe maternal demographics, pregnancy characteristics, medical risk factors, and selected delivery complications. Among these variables are five maternal morbidity indicators, which I used to create a composite indicator, coded as 1 if any of the five morbidity indicators are present and 0 otherwise.

Although these five indicators do not capture the full scope of the 21 SMM indicators defined by the CDC, this dataset is the only publicly available, nationally standardized measure of severe maternal morbidity at this scale, making it a strong option

for population-level modeling. The indicators reflect clinically severe, validated outcomes maintained and disseminated by the CDC that serve as a reasonable proxy for severe maternal morbidity in the data (NCHS, 2025b). While the analysis aims to predict SMM, the rarity of the outcome and the nature of the data limit the feasibility of producing a highly predictive model. Instead, the goal is to compare machine learning approaches and examine what different techniques can— and cannot— reveal about SMM.

The NVSS data file contains a row for each registered birth in the United States in 2024, or about 3.6 million rows. Given the length of the data, it was imperative to limit width by creating a trimmed version of the dataset, retaining only the variables needed for analysis. In addition to the SMM outcome, I kept the 11 predefined clinical and obstetric risk factors included by NVSS. These range from eclampsia to Previous Preterm Birth, to a history of cesarean delivery. I also included a composite binary indicator for the presence of any one of these risk factors. In addition to risk factors, I retained socioeconomic proxies, maternal demographics, pregnancy context, and health behaviors in the trimmed dataset. For a full list of variables used in the analysis, see Appendix Table 1.

After trimming the file, I created the SMM composite as described above. To prepare the predictor variables, I recoded variables from yes, no, and unknown to 1, 0, and missing. This approach preserves uncertainty in the original data and ensures that predictors are compatible with machine learning algorithms. After recoding, I assessed missingness across all predictor variables. Two risk-factor variables— use of assistive reproductive technology and fertility enhancing drugs— were missing information for about 70% of the observations, so I dropped from the dataset. An additional eight risk-factor variables had

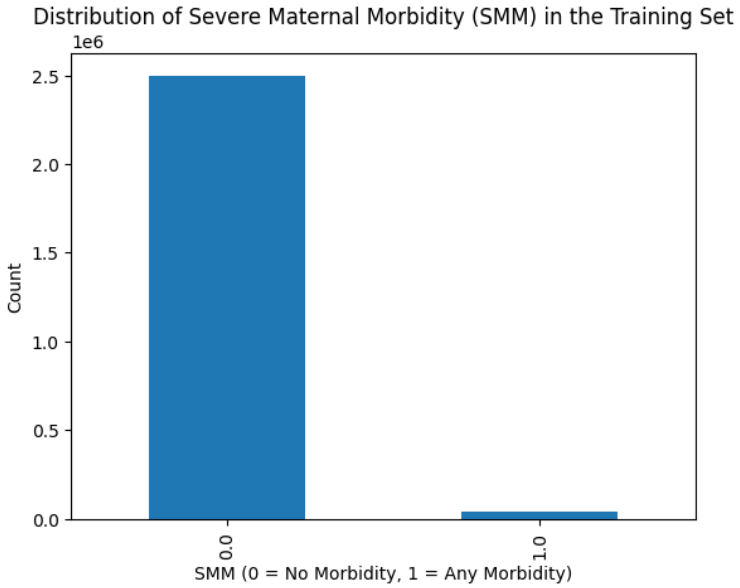
minimal missing data, with less than one percent missing, each. For these variables, I imputed the modal value, effectively assigning “no risk” to a small number of observations, given the relative rarity of the conditions captured by these indicators.

In preparation for modeling, I also encoded categorical variables. I used one-hot encoding for nominal variables, and ordinal encoding for ordinal variables in order to preserve natural ordering. Additionally, I scaled continuous variables so that numeric features were on a comparable scale. After completing data preprocessing, I split the cleaned, trimmed dataset into training and test sets using a 70/30 split and stratified sampling to preserve the low prevalence of SMM in both subsets.

In addition to the challenge of the size of the dataset, the relative rarity of SMM became evident during the cleaning process. Comprising only 1.53% of cases in both the train and the test sets, the extreme class imbalance of the target variable emerged as a central feature of the data and influenced subsequent modeling decisions.

Most predictor variables in the dataset are binary, and many are imbalanced, which limits how much could be learned from univariate visualizations. Visualizing the distribution of SMM confirmed the extreme class imbalance, reinforcing that subsequent analyses would need to focus on how modeling approaches handle rare outcomes. Figure 1 displays the distribution of SMM in the training set.

Figure 1. Distribution of SMM in the Training Set



Supervised Learning

To examine the data using supervised learning, I opted to use logistic regression and random forest. Logistic regression models the probability of a categorical outcome— in this case SMM versus no SMM— by estimating a linear relationship between predictors and the log-odds of the outcome. Logistic regression relies on a fixed decision threshold, so it tends to favor the majority class when the outcome is rare, unless class weights or thresholds are adjusted. This pattern is evident in the default model's failure to detect a single case of Severe Maternal Morbidity (see Table 1).

Given the rarity of SMM, model evaluation requires metrics beyond accuracy. The default logistic regression model achieved an accuracy of 98.5% on both the training and test sets by predicting "no SMM" for every observation. While that prediction was indeed accurate 98.5% of the time, the model provides no meaningful insight into performance on

the outcome of interest, SMM. Rather than rely on accuracy, I focused on precision, recall, and the F1 score for the positive class when evaluating model performance.

To improve logistic regression performance on the minority class, I tuned the model by adjusting the regularization strength and the class-weight parameter. Adjusting regularization strength limits the influence of individual predictors and reduces overfitting, while class weights penalize misclassification of SMM more heavily than misclassification of non-SMM. I used cross-validated F1 for the positive class to select tuning parameters, prioritizing balanced performance between precision and recall above accuracy.

Introducing balanced class weights shifted the model away from always predicting the majority class and increased sensitivity to SMM. Compared to the default, which had zero recall, the tuned model achieved a recall of 0.60 for SMM on both the training and test sets. This improvement comes at the cost of accuracy, which declined substantially to 0.60; and precision, which remained low at 0.02, resulting in an F1 score of 0.04. While the tuned model identified substantially more SMM cases than the default model, it did so by generating a large number of false positives, illustrating the trade-off between recall and precision when modeling rare outcomes.

Table 1. Logistic Regression Performance on SMM (Positive Class Metrics)

	Dataset	Accuracy	Precision	Recall	F1
Default Model	Test	.985	0.00	0.00	0.00
Tuned- Best Model	Train	.60	0.02	0.60	0.04
	Test	.61	0.02	0.60	0.04

Random forest models build many individual decision trees on random subsets of data, then aggregate their predictions to reduce variance, improve stability, and capture

non-linear relationships. I used a random forest to examine whether a flexible, non-linear model could identify structure in the data beyond that captured by logistic regression.

As with logistic regression, the default random forest model failed to meaningfully detect SMM cases. Although the default parameters achieved high overall accuracy, recall for SMM was near zero, indicating that the model primarily predicted the majority class, and provided little insight into performance on the rare outcome of interest (see Table 2). To improve performance, I tuned the random forest by adjusting the number of trees, maximum tree depth, and the number of predictors considered at each split. Given the size of the dataset, I conducted hyperparameter tuning on a stratified 20 percent subsample to preserve class balance while keeping the grid search computationally feasible.

The best-performing random forest model had the most flexible hyperparameters, with unrestricted tree depth and logarithmic feature sampling. This model improved detection of SMM in the training data, but these gains did not generalize. On the test set, both precision and recall for SMM returned to zero, indicating substantial overfitting. While the tuned random forest occasionally predicted SMM correctly, it failed to do so consistently, highlighting the limits of highly flexible models when applied to extremely imbalanced outcomes.

Table 2. Random Forest Performance on SMM (Positive Class Metrics)

	Dataset	Accuracy	Precision	Recall	F1
Default Model	Test	.983	0.07	0.01	0.01
Tuned- Best Model	Train	.99	0.56	0.12	0.19
	Test	.98	0.04	0.00	0.01

Notably, the two supervised models responded to tuning in fundamentally different ways. In logistic regression, adding balanced class weights increased recall by encouraging the model to predict SMM more frequently. However, the resultant and substantial increase in false positives led to very low precision. In contrast, the tuned random forest became more conservative. Some SMM predictions were correct, leading to an increase in precision, but the model failed to predict reliably, and recall dropped to zero on the test set, underscoring the limited utility for identifying rare cases.

Primary Component Analysis (PCA) and Unsupervised Learning

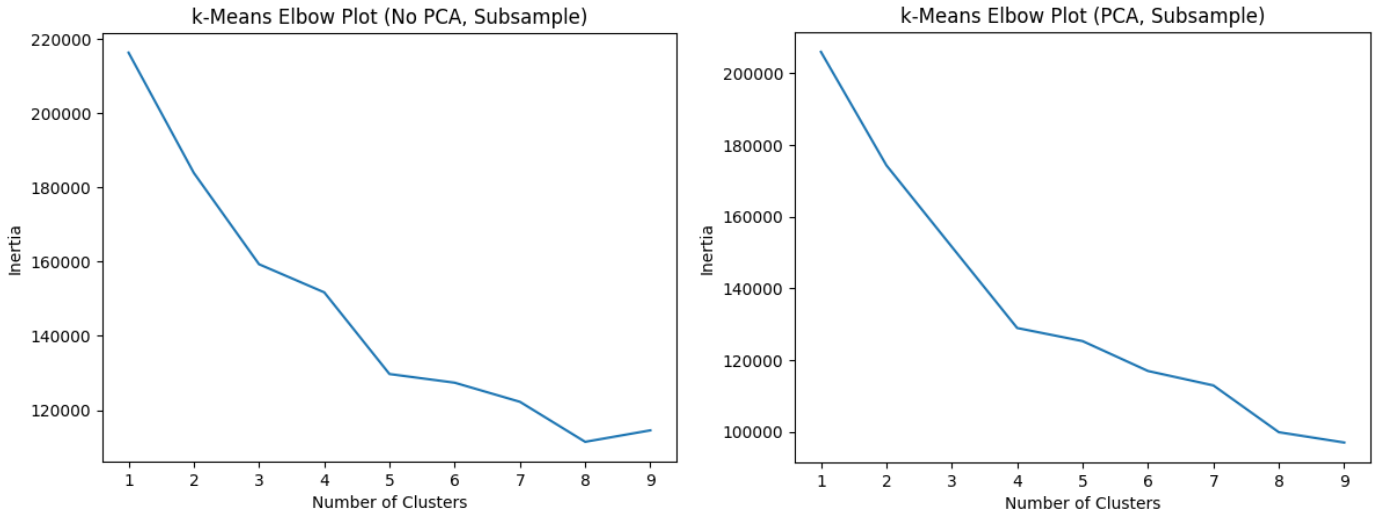
Prior to applying unsupervised learning methods, I ran PCA on a 10,000-row subsample of the training data and found that 16 components are required to explain 95 percent of the variance in the predictors. To assess whether dimensionality reduction improved model performance, I refit the tuned logistic regression model using only these 16 components; however, performance did not meaningfully improve. SMM recall increased modestly from 0.60 to 0.64, but overall performance remained largely unchanged and continued to be driven by the majority class.

Table 3. Logistic Regression Performance With and Without PCA

	Dataset	Class	Precision	Recall	F1
Best Model	Train	No SMM	0.99	0.60	0.75
		SMM	0.02	0.60	0.04
	Test	No SMM	0.99	0.61	0.75
		SMM	0.02	0.60	0.04
Best Model, PCA	Train	No SMM	0.99	0.54	0.70
		SMM	0.02	0.64	0.04
	Test	No SMM	0.99	0.54	0.70
		SMM	0.02	0.64	0.04

To explore whether unsupervised learning could uncover meaningful structure in the data, I applied three clustering algorithms— k-means, agglomerative clustering, and DBSCAN— both with and without PCA as a pre-processing step. K-means partitions observations into a fixed number of clusters by minimizing variance within clusters, agglomerative clustering begins with each data point as its own cluster and iteratively merges similar observations to build clusters hierarchically, and DBSCAN identifies clusters based on local density rather than a fixed number of groups and classifies points in low-density areas as noise. For k-means, I used elbow plots to guide cluster selection. Without PCA, the elbow plot was not strongly conclusive, but I selected three clusters as a reasonable compromise. After PCA, the elbow plot suggested four clusters.

Figure 2. Elbow Plots With and Without PCA



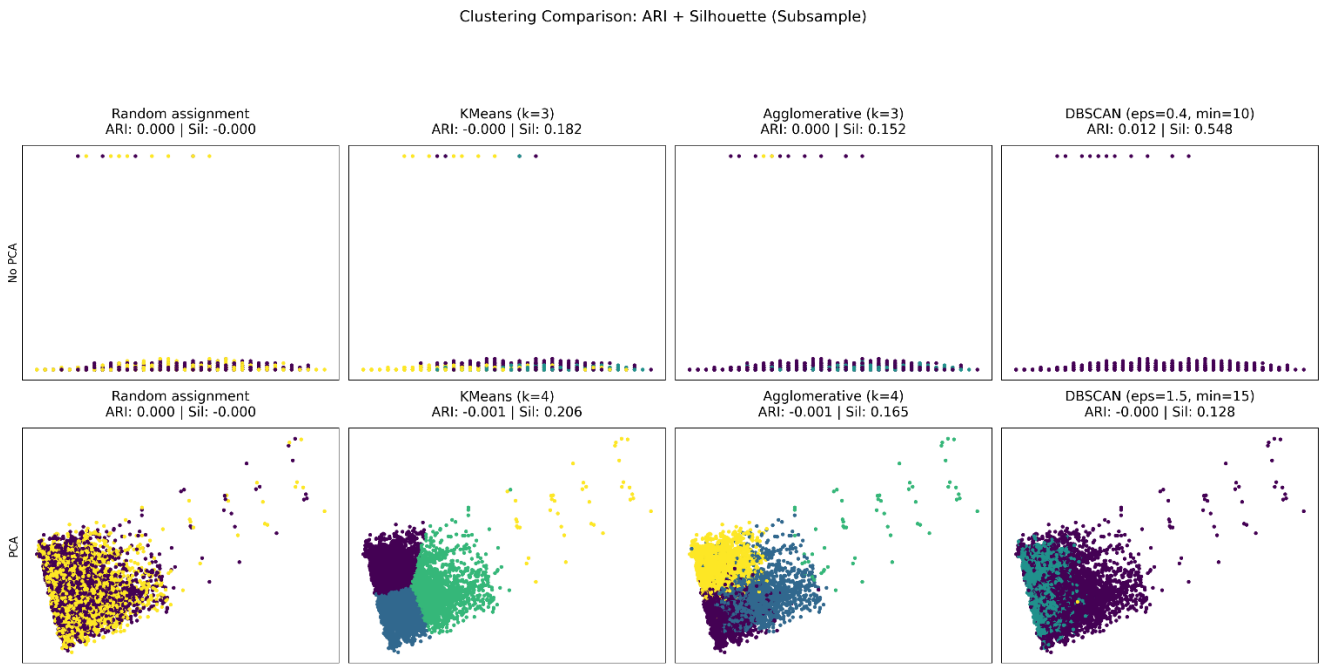
Visual inspection of the clusters across all three methods shows that PCA improves the geometric separation of clusters (see Figure 3). Without PCA, clusters appear compressed along a single axis, while the reduced feature space produces more visually distinct and interpretable clusters, though some overlap remains. This change suggests

that PCA helped organize the data in a lower-dimensional space, but did not eliminate ambiguity between clusters.

Despite improved visualization, clustering performance did not meaningfully improve after applying PCA. Adjusted Rand Index (ARI) values remain near zero, indicating that cluster assignments were no better than random, and suggesting that the variance captured by the principal components is not strongly associated with SMM (see Figure 3).

Silhouette scores offer a similar conclusion. PCA produced slight increases in silhouette scores for both k-means and agglomerative clustering, indicating marginal improvement in cluster separation, but the scores remained weak overall. DBSCAN yielded negative silhouette values regardless of PCA, indicating poor density-based separation (see Figure 3). Together, these results suggest that while PCA modestly improved cluster visualization, it was not sufficient to produce substantively meaningful clusters in the data.

Figure 3. Clusters, ARI, and Silhouette With and Without PCA



Together these projects offered the opportunity to apply classroom learning to a real-world phenomenon. From a pedagogical standpoint, the projects were effective: adjusting elements such as class weighting, model flexibility, and evaluation metrics made the trade-offs between recall, precision, and overall performance concrete. However, despite tuning and preprocessing, the models exhibited little practical predictive power for severe maternal morbidity.

If I were to revisit this work, I would explore some of the modeling techniques that my classmates with imbalanced datasets found effective, such as XGBoost. I would also examine the clustering structure more deliberately. Rather than treating the number of clusters as an outcome in itself, I would like to explore whether the apparent three and four cluster solutions reflect meaningful subgroups in the data, or whether they reflect modeling and preprocessing choices rather than substantive differences in the data. Finally, I would examine other data sources that sacrifice breadth for greater clinical detail, which may be better suited for modeling rare and complex outcomes like SMM.

Appendix

Table 1. Codebook of Variables Used in Analysis

Variable Name	Description
Maternal morbidity indicators	
mm_mtr	Maternal Transfusion
mm_plac	Perineal Laceration
mm_rupt	Ruptured Uterus
mm_uhyst	Unplanned Hysterectomy
mm_aicu	Admit to Intensive Care
SMM*	Maternal Morbidity Composite, Calculated
Maternal risk factors	
rf_pdiab	Pre-pregnancy Diabetes
rf_gdiab	Gestational Diabetes
rf_phype	Pre-pregnancy Hypertension
rf_ghype	Gestational Hypertension
rf_ehype	Hypertension Eclampsia
rf_ppterm	Previous Preterm Birth
rf_inftr	Infertility Treatment Used
rf_fedrg	Fertility Enhancing Drugs
rf_artec	Asst. Reproductive Technology
rf_cesar	Previous Cesarean
rf_cesarn	Number of Previous Cesarean
no_risks	No Risk Factors Reported
Socioeconomic proxies	
meduc	Mother's Education

WIC	Participation in the Women, Infants, and Children (WIC) nutrition program
pay_rec	Payment Source for Delivery Recode
Race and ethnicity	
mracehisp	Mother's Race/Hispanic Origin
mhisp_r	Mother's Hispanic Origin Recode
mrace6	Mother's Race Recode, 6 categories
Health behaviors and prenatal care	
bmi_r	Maternal pre-pregnancy body mass index (BMI), recoded
cig0_r	Cigarettes Before Pregnancy Recode
precare5	Month Prenatal Care Began Recode
Demographics and pregnancy history	
mager	Mother's Single Years of Age
dplural	Plurality Recode
lbo_rec	Live Birth Order Recode
tbo_rec	Total Birth Order Recode
priorlive	Prior Births Now Living
priordead	Prior Births Now Dead
priorterm	Prior Other Terminations
Gestational age	
gestrec10	Combined Gestation Recode 10

* Constructed composite indicator of severe maternal morbidity, coded as 1 if any maternal morbidity indicator is present. All other variables pulled directly from NCHS, 2025a.

Sources

Centers for Disease Control and Prevention. (2021). Severe maternal morbidity in the United States. U.S. Department of Health & Human Services. https://www.cdc.gov/reproductivehealth/maternalinfanthealth/severematernalmorbidity.html

National Center for Health Statistics. (2025). National Vital Statistics System natality data (2024) [Public-use dataset]. National Bureau of Economic Research. https://data.nber.org/nvss/natality/dta/

National Center for Health Statistics. (2025). User guide to the 2024 Natality public use file [PDF]. Centers for Disease Control and Prevention. https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/DVS/natality/UserGuide2024.pdf