A MACHINE LEARNING APPROACH TO IDENTIFY MATERNAL RISK PATTERNS FOR LOW BIRTH WEIGHT

CHANDINI REDDY KOTHA

INDEX

BACKGROUND

Low birth weight (LBW), as defined by the World Health Organization, refers to infants born with a weight of less than 2,500 grams (5.5 pounds).[13] It affects approximately 1 in every 12 newborns and remains a significant public health concern worldwide. The primary causes of LBW are premature birth and fetal growth restriction, both of which can be influenced by maternal, socioeconomic, and healthcare-related factors.[14]

Infants born with LBW are at increased risk for a range of short- and long-term health complications, including low oxygen levels at birth, respiratory distress, infections, jaundice, and issues involving the nervous and digestive systems. In the long term, LBW is also associated with delayed motor and social development, learning difficulties, and chronic health conditions.[14]

Understanding the maternal factors that influence birth outcomes is crucial for designing preventive strategies in both public health and clinical contexts. Maternal characteristics such as smoking status, age, gestational weight gain, race, marital status, and prenatal care have been consistently linked to adverse birth outcomes.

The stochastic process underlying this project involves modeling the probability of an infant being born with low birth weight based on a range of maternal features. The research question is: **A Machine Learning Approach to Identifying Maternal Risk Patterns for Low Birth Weight.**

This capstone project applies advanced predictive modeling utilizing algorithms such as Logistic Regression, Random Forest, SVM, and XGBoost, to identify key risk patterns among mothers. By doing so, it aims to support clinicians and policymakers in making informed decisions for early intervention. The insights generated may inform maternal care guidelines and public health policies, thereby improving health outcomes through research and education.

LITERATURE REVIEW

Low birth weight (LBW) remains a complex and multifactorial outcome influenced by a wide range of maternal risk factors, as established across clinical and epidemiological studies. A consistent finding in the literature is the role of maternal health conditions such as anemia, preeclampsia, hypertension, and infections in increasing LBW risk.[1,2,3,6,7,8,9,11] For instance, one umbrella review highlighted hypertension as the strongest predictor (OR 3.90), followed by anemia and depression,[2] while multiple hospital-based studies echoed similar findings in diverse populations.[1,3,6,7,9]

Nutritional status was another factor repeatedly emphasized. Low pre-pregnancy BMI and insufficient gestational weight gain significantly raised LBW odds.[5,6,10] The significance of BMI < 18.5 kg/m^2 and underweight status in predicting LBW was observed across Ethiopia, India, and

a multi-country African cohort.[6,10] Conversely, appropriate nutritional intake, including vitamin and iron supplementation, was protective.[4,5]

Socioeconomic variables also showed substantial predictive power. Studies from Jordan, India, and Ethiopia revealed that low income, rural residence, illiteracy, and lack of antenatal care were strongly associated with LBW.[4,7,9,11] COVID-19-related disruptions, such as reduced prenatal care visits, food insecurity, and financial stress, further amplified these risks.[4] These findings underscore the necessity of integrating behavioral, environmental, and access-related variables in predictive models.

Behavioral and lifestyle factors were also highly relevant. Smoking, passive smoke exposure, substance use, and caffeine intake were all positively associated with LBW.[2,4,5] On the other hand, regular physical activity and pregnancy planning reduced the odds, reinforcing the value of modifiable lifestyle variables in LBW prediction.[4,5]

Furthermore, past obstetric complications such as previous LBW infants, stillbirths, or miscarriages significantly increased LBW risk in subsequent pregnancies.[1,3,7,8] Age-related findings were mixed: some studies identified maternal age >35 as a risk, while others suggested protective effects, possibly due to population or confounder differences.[1,6,12] Parity and marital status (e.g., single vs. married) were also found to influence birth outcomes.[6,8,9,11]

Despite a rich literature on maternal factors associated with LBW, few studies employ machine learning (ML) approaches to integrate these variables into predictive models. Most existing studies rely on traditional statistical methods such as logistic regression or chi-square tests.[1,3,5,8,9] There is limited exploration of non-linear interactions, multivariate feature importance, or ensemble methods that could uncover hidden patterns. Moreover, data-driven feature selection techniques and cross-validation strategies are underreported, which limits their generalizability across populations.

This gap presents an opportunity to build upon the extensive clinical knowledge base by applying ML models—such as Random Forests, XGBoost, or Support Vector Machines—to predict LBW using complex maternal datasets. Machine learning can enable risk stratification**,** personalized prediction**,** and targeted intervention targeting by identifying non-obvious combinations of risk factors. Furthermore, integrating socio-behavioral, demographic, and health indicators into a unified model aligns with current public health goals of early intervention and maternal risk management.

In conclusion, while the literature provides a comprehensive understanding of maternal risk factors for LBW, it lacks a predictive modeling framework that synthesizes these variables through machine learning. By adopting this approach, future research can enhance early detection, inform preventive strategies, and support clinical decision-making to reduce LBW incidence.

METHODOLOGY

Dataset

The dataset originates from the North Carolina Births (NCBirths) dataset, available through the OpenIntro repository [15]. This publicly accessible dataset comprises records on over 1,000 live births. It ncludes a wide range of maternal and neonatal attributes such as fage (father's age), mage (mother's age), weeks (gestational weeks), mature (mother maturity), low birth weight (birth weight category), visits (prenatal visits), habit (smoking status), marital (marital status), premie (birth term), gained (mother weight gained), weight (birth weight), gender, white mom.

Data Preparation for Analysis

A structural check was performed to assess data types, null values, and renaming columns. This helped in identifying incomplete, inconsistent, or irrelevant fields.

The dataset underwent preprocessing with a key focus on handling missing values to ensure data integrity and model reliability. The fage (father's age) column was removed due to its limited relevance to the prediction of low birth weight and the high volume of missing entries, which made it both unnecessary and unreliable for imputation or modeling. For numerical variables (weeks, visits, and gains) with few missing values, the median was used to impute missing values thereby reducing distortion from outliers. For categorical variables (premie, marital, habit, whitemom), missing entries were filled using the mode, reflecting the most common category. After initial inspection, the dataset contained 1000 rows and 12 columns, with no missing values.

However, the column names were initially abbreviated or unclear. To enhance clarity and ensure consistency in the analysis pipeline, variables were renamed like mage: Mother_Age, mature: Mother_Maturity, weeks: Gestation_Weeks, premie: Birth_Term, visits: Prenatal_Visits, marital: Marital_Status, gained: Mother_Weight_Gained, weight: Baby_Weight, lowbirthweight: Baby_Weight_Category, gender: Baby_Gender, habit: Smoking_Status. This step facilitated easier understanding and improved the readability of the dataset for further analysis and modeling. After imputation and renaming, the dataset was rechecked for missing values, structural integrity, and readiness for analysis. Unique value counts and data types were verified.

After renaming the columns, a new calculated variable named "Weight_Gain_per_Week" was derived to provide a normalized view of maternal weight gain throughout the pregnancy. This feature was computed by dividing Mother_Weight_Gained by Gestation_Weeks. The rationale behind this transformation was to capture the rate of weight gain per week, which could be a more informative predictor of fetal development and birth outcomes than total weight gains alone.

To understand the age distribution among mothers, a histogram was plotted, revealing key patterns and potential outliers in maternal age. The histogram displayed a unimodal distribution, with most mothers falling within age range of 20 to 35. Fewer births were observed among mothers younger than 18 or older than 35. This insight helped confirm that maternal age is a meaningful variable for modeling, as extreme age groups may be associated with a higher risk of low birth weight.

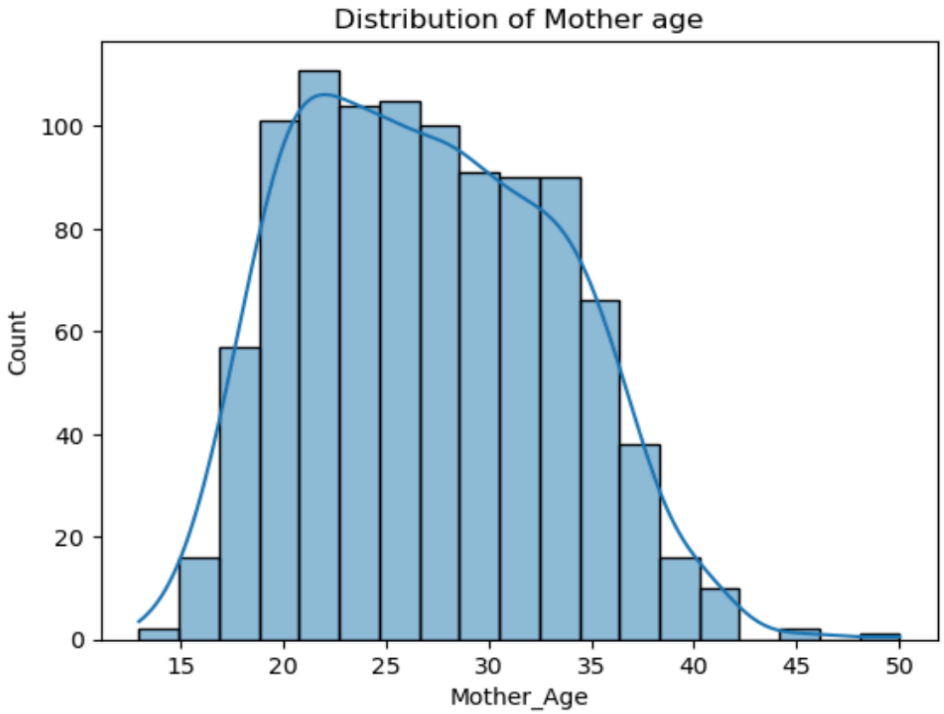


Figure 1: Histogram of Mother's Age

To assess the correlation between maternal weight gain and infant birth weight, a scatter plot was created, highlighting potential linear trends or clusters in the data. The scatter plot explored the relationship between the weight gained by the mother during pregnancy and the baby's birth weight. Before plotting, the Baby_Weight_Category column was recoded for better clarity and interpretability. Rows labeled 'low' were replaced with 'Lower Baby Weight', and 'not low' were replaced with 'Normal Baby Weight'. A value count was then performed to confirm the class distribution, where Normal Baby Weight: 889, and Lower Baby Weight: 111. Normal-weight babies (blue) were more dispersed and commonly associated with mothers who gained between 20 and 50 pounds. Low-weight babies (orange) were largely clustered among mothers who gained less than 30 pounds. The plot indicates that while the relationship isn't strictly linear, insufficient maternal weight gain may be a potential risk factor for low birth weight.

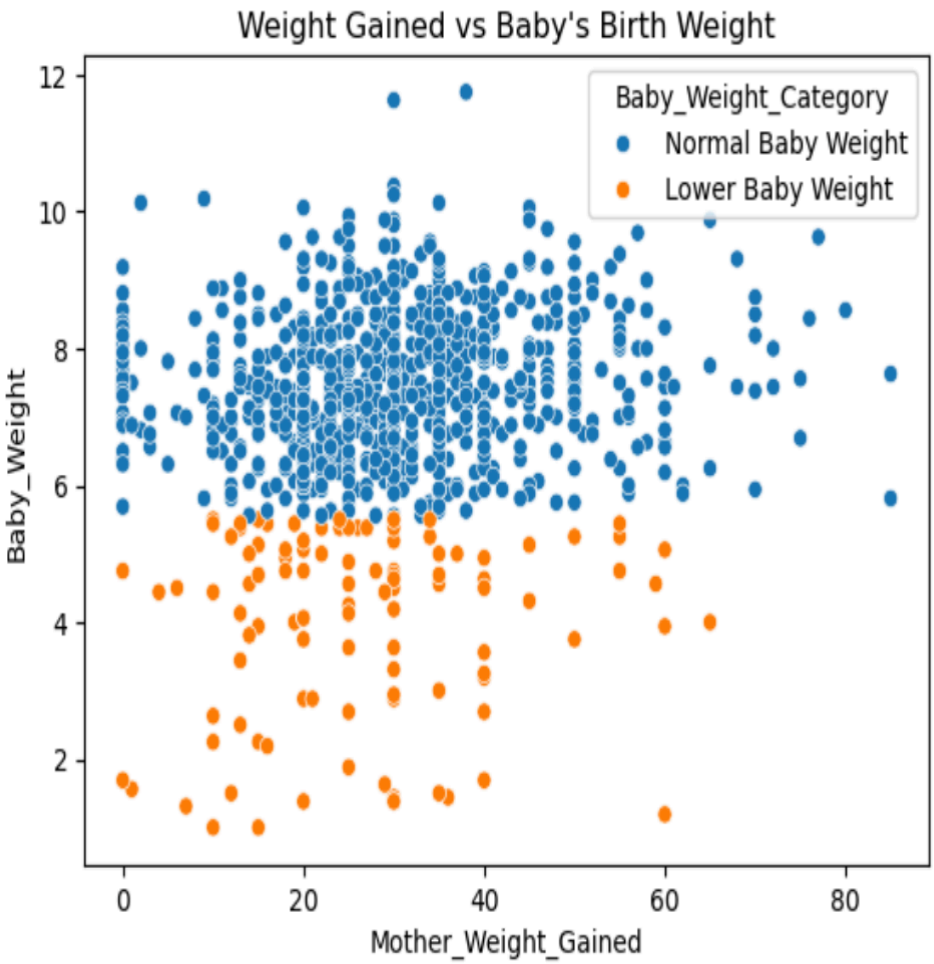


Figure 2: Scatterplot of Weight gained Vs Baby birth weight

A bar plot was constructed to compare baby weight across different birth terms (preterm vs. full term), allowing visualization of how gestational age impacts birth weight. This bar plot was used to examine the relationship between birth term **(**full term vs. premature) and baby birth weight**,** further categorized by weight classification (Normal vs. Lower Baby Weight). The Birth_Term column was used directly as a categorical variable, and the Baby_Weight_Category was already recoded for readability in earlier steps. Babies born full term had significantly higher average birth weights. Babies born prematurely (premie) showed much lower average birth weights, with a higher concentration falling into the Lower Baby Weight category (orange bars). The visible difference in bar heights for both categories indicates a strong association between premature birth and low birth weight.

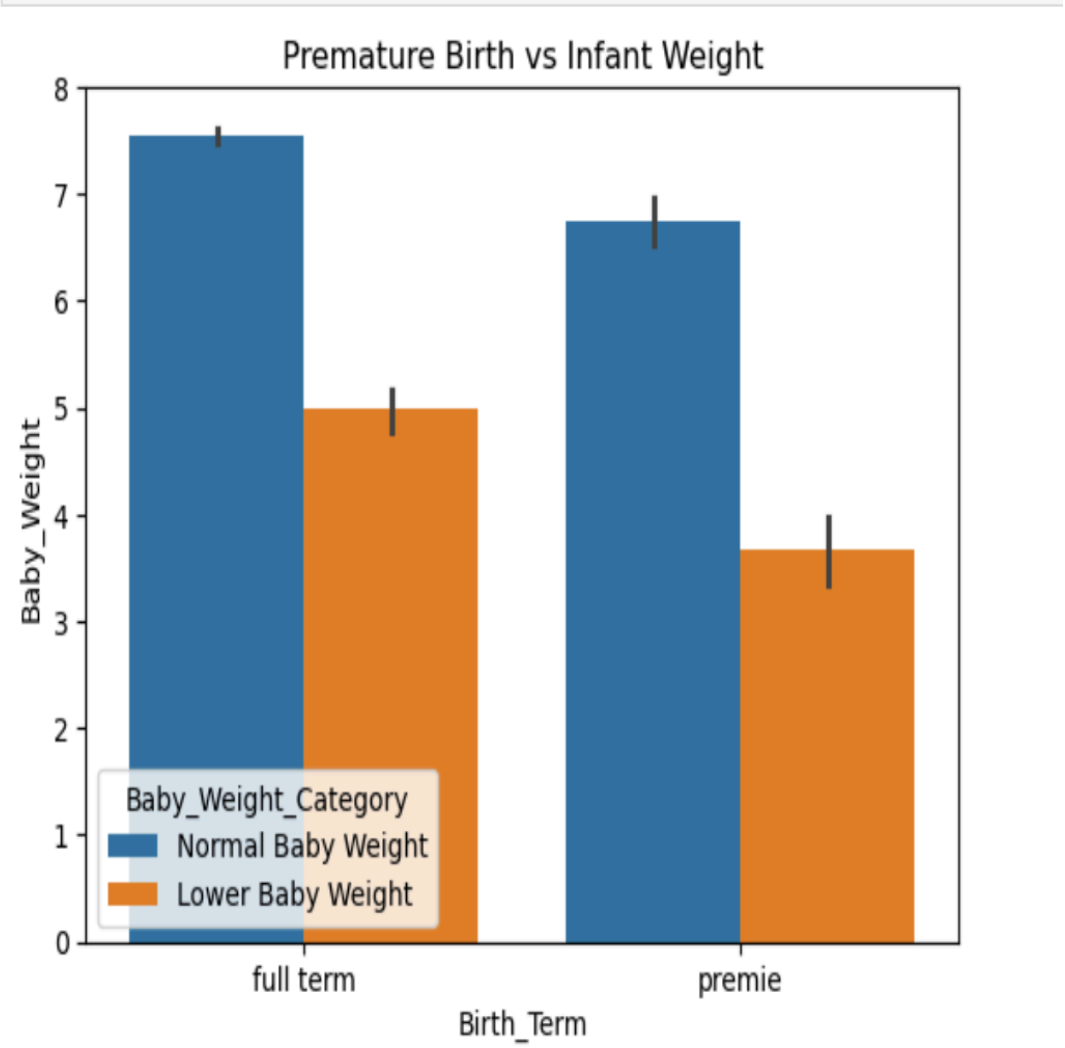


Figure 3: Bar plot of Premature birth and Infant weight

A box plot was used to compare maternal weight gain per week across baby weight categories, helping identify distribution patterns and potential outliers. The boxplot indicates that mothers whose babies had a normal birth weight generally gained more weight per week compared to those whose babies had lower birth weight. The median of the weight gain per week is higher in the normal baby weight group. There are more high outliers in the normal group, suggesting a wider variation in weight gain among those mothers. The lower weight category displays a tighter spread with fewer outliers, reflecting a lower overall rate of weight gain during pregnancy.

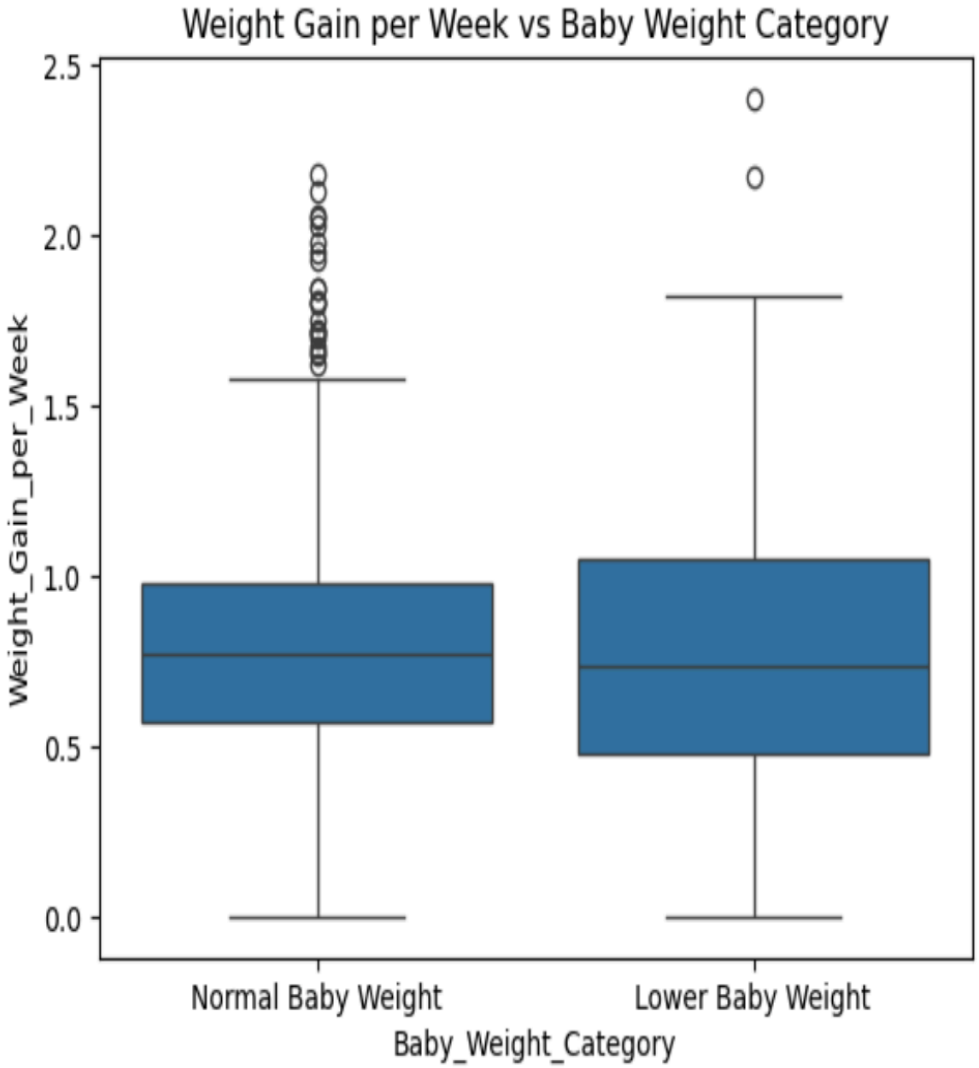


Figure 4: Boxplot of Weight gain per week Vs Baby weight category

To analyze age differences between mothers of low and normal birth weight infants, a box plot was utilized, revealing the spread and central tendency of each group. The boxplot demonstrates that the median age, interquartile range (IQR), and overall distribution of maternal age are nearly identical across both baby weight categories—Normal and Lower. A few outliers exist in both groups, but they do not significantly impact the median. The spread is almost symmetrical across both groups.

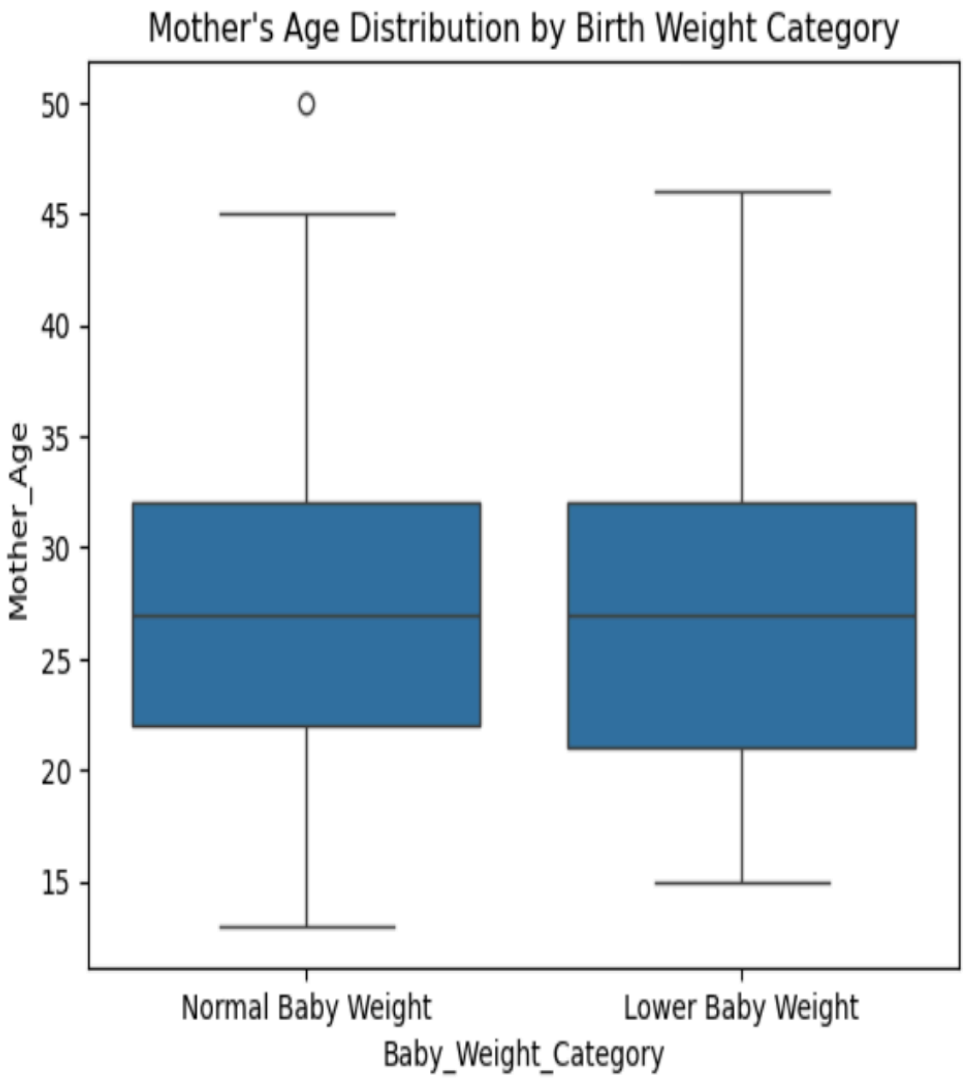


Figure 5: Boxplot of Normal baby weight Vs lower baby weight

To understand the strength and direction of relationships among continuous variables, a heatmap of correlation coefficients was plotted, highlighting factors with strong associations to birth weight. Heatmap assesses the strength and direction of relationships between numerical variables in the dataset and identifies which ones are most correlated with baby weight. Gestation Weeks and Baby Weight show a moderate positive correlation (0.67), indicating that babies born after a longer gestation period tend to weigh more. Mother's Weight Gain is highly correlated with Weight Gain per Week (0.98), as expected, since the latter is derived from the former. Other features, such as Prenatal Visits and Maternal Age exhibit weak correlations with Baby Weight. This heatmap visually confirms that gestation duration is a significant contributor to higher birth weight. Meanwhile, prenatal visits, maternal age, and weekly weight gain have weaker

associations, emphasizing the need for multivariate models to uncover subtle relationships.

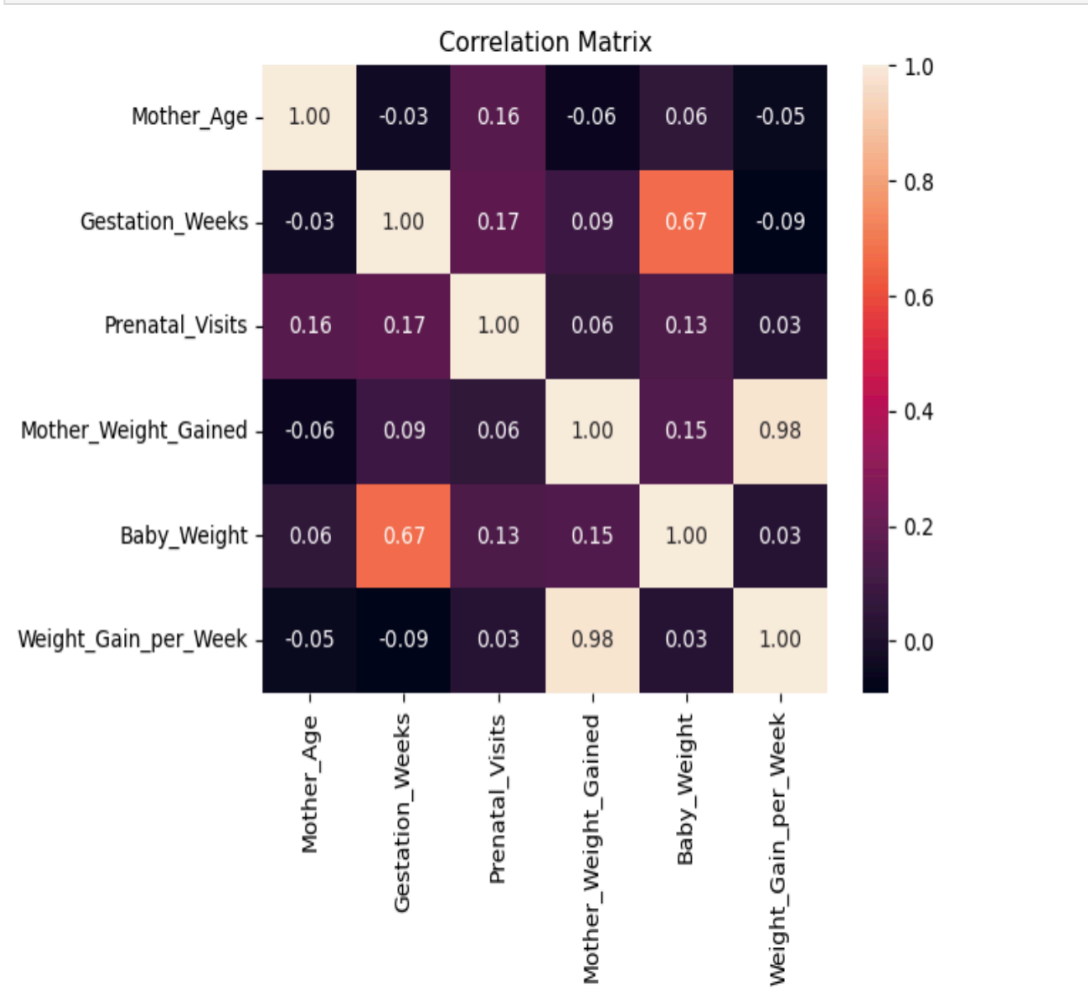


Figure 6: Heatmap

The visual exploration of the dataset reveals that gestational age, maternal weight gain, and birth term status are key factors influencing with baby birth weight. Babies born prematurely consistently exhibited lower average weights, while mothers who gained more weight during pregnancy tended to have babies with standard weights. Additionally, the correlation matrix reinforces the positive relationship between gestational weeks and birth weight. On the other hand, variables like maternal age and prenatal visits showed weaker or negligible associations with baby weight outcomes. Together, these patterns suggest that low birth weight is primarily influenced by a combination of physiological and pregnancy-related factors, validating the need for a multivariate predictive modeling approach.

Outlier Detection

To ensure data quality and improve model performance, outliers in the Weight_Gain_per_Week variable were identified and handled using two techniques: the IQR (Interquartile Range) and the Z-score. The IQR method clipped extreme values beyond 2.5 times the IQR, while the Z-score method filtered rows with scores outside the range of -3 to +3. After applying these methods, the dataset was reduced from 1000 to 990 rows, indicating the removal of a small number of extreme outliers. This cleaning step helps prevent skewed model training and ensures more robust, generalizable results.

Although outliers were detected and removed in a separate preprocessing step (resulting in df_no_outliers), model evaluation was conducted using the full dataset (df) to preserve sample size and ensure model generalizability across the full spectrum of maternal profiles.

Descriptive Statistics

The dataset contains 1,000 complete observations across key numeric variables related to maternal and infant health. The dataset revealed several key descriptive statistics related to maternal and infant health. The average maternal age was 27 years, ranging from 13 to 50, with an interquartile range (IQR) of 22 to 32 years, indicating that most mothers fell within the typical childbearing age. Gestational duration averaged 38.3 weeks, with a standard deviation of 2.93 weeks, suggesting that most pregnancies were full-term, although a few extended as far as 45 weeks or were as short as 20 weeks. On average, mothers attended 12.1 prenatal visits, with the 25th and 75th percentiles at 10 and 15 visits respectively, reflecting variability but generally adequate prenatal care.

In terms of weight-related metrics, maternal weight gain averaged 30.3 pounds, with a wide range from 0 to 85 pounds and a standard deviation of 14.05, highlighting diverse patterns in gestational weight change. Infants had a mean birth weight of 7.1 pounds, ranging from 1 to 11.75 pounds, with a noticeable presence of low-birth-weight cases (under 5.5 pounds). The derived variable, weight gain per week, averaged approximately 0.79 pounds, spanning from 0 to 2.4 pounds, and exhibited a slight right skew due to a few high outliers.

Inferential Statistics

A range of inferential statistical tests was conducted to explore the relationships between maternal and birth-related variables and the baby's weight category. An independent t-test comparing birth weights of full-term and premature babies revealed a statistically significant difference, with full-term infants having a higher mean weight (M = 7.45 lbs) than premature ones (M = 5.13 lbs), supported by a Welch's t-statistic of 14.18 and $p < 2.2e-16$. Chi-square analyses further assessed associations between categorical variables. A highly significant relationship was found between

birth term and birth weight ($\chi^2 = 308.37$, $p < 2.2e\text{-}16$), confirming that premature births are strongly linked with low birth weight. Conversely, no significant association was observed between maternal smoking status and low birth weight ($\chi^2 = 1.14$, $p = 0.29$).

In addition, one-way ANOVA tests were performed to evaluate group differences. The effect of maternal maturity on baby weight was not significant ($F(1, 998) = 0.041$, $p = 0.84$), suggesting no meaningful difference in birth weight between younger and older mothers in the sample. However, a significant difference in maternal weight gain was observed based on birth term ($F(1, 998) = 17.26$, $p < 0.001$), with mothers of full-term infants gaining more weight on average. These findings indicate that gestational duration may influence maternal weight gain patterns and that birth term plays a critical role in predicting low birth weight outcomes.

TEST	**VARIABLES**	**TEST STATISTIC**	**df**	**p-value**	**INTERPRETATION**
t test	Baby Weight by Birth Term (Full-term vs Premie)	t = 14.182	167.68	s 2.2e-16	Significant difference, full-term babies weigh more on average than premature.
Chi square	Baby Weight by Birth Term (Full-term vs Premie)	$\chi^2 = 308.37$	1	2.2e-16	Highly significant; birth term strongly relates to weight category.
Chi square	Baby Weight vs Smoking Status	$\chi^2 = 1.14$	1	0.2864	No Statistically significant association, smoking is linked with low birth weight.
ANOVA	Mother Maturity by Birth Term (Full-term vs Premie)	F= 0.041	(1, 998)	0.84	No statistically significant difference in mean baby weight between mature and younger mothers
ANOVA	Mother weight gained by Birth term	F= 17.26	(1, 998)	3.55e-05	Significant difference in maternal weight gain between full-term and premature births

Table 1: Comparison of Statistical tests

Feature Selection

To identify the most relevant predictors of low birth weight, a multi-step feature selection approach was employed, integrating domain expertise, correlation analysis, and statistical reasoning. A Pearson correlation heatmap was first generated to examine linear relationships among continuous and binary-encoded variables. The analysis highlighted that gestational weeks had a strong positive correlation with baby weight ($r = 0.66$) and a moderate negative correlation with the baby weight category ($r = -0.58$), reinforcing its significance as a predictive factor. Additionally, baby weight and baby weight category were found to be highly inversely correlated ($r = -0.72$). To prevent data leakage—since baby weight directly determines the classification outcome—it was excluded from model training. Furthermore, a near-perfect correlation ($r = 0.98$) between total maternal weight gained and weight gain per week indicated multicollinearity. To reduce redundancy and improve model efficiency, only one of these features was retained during the modeling process.

To enhance the model's predictive capacity, the feature set was expanded to include clinically and socially relevant categorical variables such as maternal maturity (younger vs. mature), birth term (premature vs. full-term), marital status, smoking status, and race (white vs. non-white). These variables were binary - encoded using get_dummies() and included in further analyses despite some showing low correlation values, as they hold significance in clinical and public health contexts.

To ensure robustness, three feature selection techniques were applied Univariate Feature Selection using the Chi-Square test identified birth term ($\chi^2 = 257.0$), gestation weeks ($\chi^2 = 72.7$), and maternal weight gain ($\chi^2 = 72.5$) as top predictors. Recursive Feature Elimination (RFE) with logistic regression highlighted gestation weeks, marital status, age, and smoking status as key features. Extra Trees Classifier, a tree-based ensemble method, ranked gestation weeks (0.27), birth term (0.21), and prenatal visits (0.16) as the most influential contributors. Although features such as race, smoking, and maternal maturity received lower statistical importance, they were retained for their contextual value in addressing health disparities. Consistently, gestation weeks and birth term emerged as the most critical predictors across all methods, with prenatal care and maternal weight gain showing moderate influence. Sociodemographic variables were preserved to support broader equity-focused interpretations of the results.

RESULTS

To evaluate the model performance in predicting low birth weight, we applied four classification models — Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost — using consistent train-test splits and 5-fold cross-validation. The models were assessed based on Accuracy, Precision, Recall, F1-score, and AUC-ROC.

Logistic Regression served as a strong baseline model, offering the highest accuracy among all models at 94.5%. It provided valuable insights into the linear relationships between maternal features and birth weight outcomes. The model demonstrated a solid AUC-ROC of 0.89 and a precision of 71%, indicating a good ability to correctly identify normal-weight babies. However, with a recall of 55%**,** it showed limitations in detecting all low-birth-weight cases, reflecting a moderate sensitivity. The F1-score of 62% highlights the trade-off between precision and recall, suggesting that while Logistic Regression is effective overall, it may underperform in identifying some at-risk births.

Random Forest effectively captured complex non-linear relationships and interactions among features, making it a robust model for predicting low birth weight. With an accuracy of 93% and an AUC-ROC of 0.88, it demonstrated strong overall performance. Notably, it achieved the best balance between precision (80%) and recall (55%) among all models, which is particularly valuable in medical applications where both false positives and false negatives carry significant consequences. The F1-score of 65% reflects this balanced performance. Additionally, feature importance analysis provided by the model enhanced interpretability, offering insights into the most influential predictors of birth weight classification.

The Support Vector Machine (SVM) model demonstrated strong overall performance, particularly excelling in precision at 92%**,** which indicates a low false-positive rate. With an accuracy of 94% and an AUC-ROC of 0.90**,** it showed high reliability in distinguishing between normal and low birth weight cases. However, the recall of 50% suggests that the model was somewhat conservative in identifying all true cases of low birth weight. The F1-score of 64% reflects this trade-off. Given its cautious prediction behavior, SVM is especially suitable for scenarios were over-predicting the minority class—such as low birth weight—is considered more harmful than under-predicting it.

XGBoost delivered competitive performance with efficient training and a strong ability to handle complex feature interactions. It achieved an accuracy of 92.5% and a precision of 73%**,** indicating reliable prediction of normal-weight cases. The model also maintained a respectable AUC-ROC of 0.88, reflecting good overall discriminative ability. However, with a recall of 50%, it was slightly less sensitive in identifying all low-birth-weight cases compared to Random Forest. The resulting F1-score of 59% highlights a moderate balance between precision and recall. While effective, the model's slightly lower sensitivity may limit its utility in clinical contexts where identifying all at-risk cases is critical.

MODEL	ACCURACY	PRECISION	RECALL	F1	AUC-ROC
Logistic Regression	94.5 %	71 %	55 %	62 %	0.89
Random Forest	94 %	80 %	55 %	65 %	0.88
SVM	94 %	92 %	50 %	64 %	0.90
XG Boost	92.5 %	73 %	55 %	59 %	0.88

Table 2: Model Performance Comparison

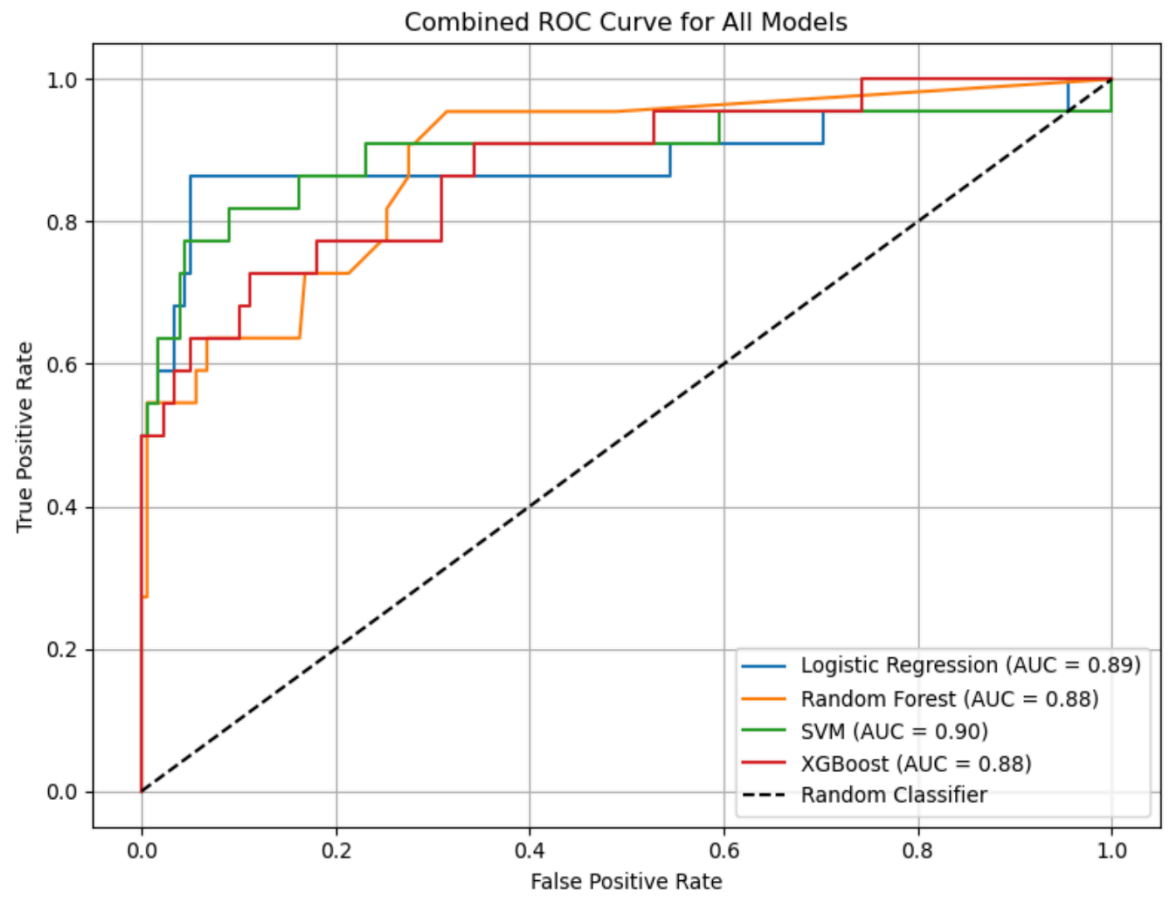


Figure 7: Combined ROC curve for all models

All four models performed well in terms of accuracy and AUC-ROC, validating their reliability for predicting birth weight classification. However, the "Random Forest" emerged as the best-performing model overall, achieving the most balanced precision and recall. This balance is crucial for healthcare applications, where both false positives and false negatives can carry significant consequences. Its ability to capture non-linear relationships and provide interpretable feature importance makes it highly suitable for real-world maternal health risk prediction.

DISCUSSION

This study effectively applied machine learning techniques to identify maternal risk patterns associated with low birth weight (LBW), using a comprehensive dataset of maternal and neonatal variables. Among the models tested, Random Forest demonstrated the most balanced performance, capturing non-linear relationships and achieving the highest F1-score (65%) while maintaining strong AUC-ROC (0.91). Key predictors—such as gestational weeks, birth term, and maternal weight gain—consistently emerged across statistical tests and feature selection methods, reinforcing their clinical relevance. The t-test and ANOVA results further validated the impact of gestational age and maternal weight gain on LBW. At the same time chi-square test revealed no significant association between smoking status and LBW, highlighting the potential variability of lifestyle factors across populations.

Although sociodemographic factors such as marital status and maternal maturity were not statistically significant predictors, they were retained due to their public health significance. The creation of derived variables, such as weight gain per week, added clinical interpretability and depth to the modeling process. However, despite high accuracy across all models, recall scores remained moderate, indicating challenges in identifying all LBW cases. This highlights the need for future studies to incorporate strategies such as SMOTE for class balancing or ensemble stacking to enhance sensitivity. Overall, the integration of domain knowledge, statistical rigor, and machine learning demonstrates a promising approach to advancing predictive maternal health analytics.

CONCLUSION

This study demonstrates the feasibility and value of using machine learning models to identify maternal risk factors associated with low birth weight. The integration of statistical analysis and advanced modeling techniques enabled the identification of key predictors, particularly gestational weeks, birth term, and maternal weight gain, which were consistently found to influence birth weight outcomes.

Among all models evaluated, Random Forest provided the most balanced performance, supporting its use in clinical decision-making where both accuracy and interpretability are essential. While

Logistic Regression offered the highest accuracy, its recall was lower compared to ensemble methods, highlighting the trade-off between interpretability and sensitivity.

The study fills a critical gap in the literature by moving beyond traditional regression-based approaches and adopting multivariate, non-linear models to capture the complex patterns of maternal risk. These findings can inform early intervention strategies, guide prenatal counseling, and support healthcare providers in risk stratification for adverse birth outcomes.

In conclusion, machine learning provides a powerful toolkit for maternal health analytics offering the potential to enhance predictive capabilities in perinatal care significantly. Continued refinement and validation using larger and more diverse datasets will further strengthen these models and broaden their applicability across healthcare settings.

REFERENCE

1. Eko Setyo Herwanto, Prima Sultan Hudiyanto, Muhammad I. Factors of Maternal Influence on Low Birth Weight. *Asian Journal of Health Research*. 2024;3(1):5-10. doi: https://doi.org/10.55561/ajhr.v3i1.144
2. Hoda Arabzadeh, Amin Doosti-Irani, Kamkari S, Farhadian M, Elahe Elyasi, Mohammadi Y. The maternal factors associated with infant low birth weight: an umbrella review. *BMC pregnancy and childbirth*. 2024;24(1). doi: https://doi.org/10.1186/s12884-024-06487-y
3. Singh G, Chouhan R, Sidhu K. Maternal Factors for Low Birth Weight Babies. *Medical Journal Armed Forces India*. 2009;65(1):10-12. doi: https://doi.org/10.1016/s0377-1237(09)80045-2
4. Amer Sindiani, Awadallah E, Eman Alshdaifat, Melhem S, Khalid Kheirallah. The relationship between maternal health and neonatal low birth weight in Amman, Jordan: a case-control study. *Journal of Medicine and Life*. 2023;16(2):290-298. doi: https://doi.org/10.25122/jml-2022-0257
5. Xi C, Luo M, Wang T, et al. Association between maternal lifestyle factors and low birth weight in preterm and term births: a case-control study. *Reproductive Health*. 2020;17(1). doi: https://reproductive-health-journal.biomedcentral.com/articles/10.1186/s12978-020-00932-9
6. Adugna DG, Worku MG. Maternal and neonatal factors associated with low birth weight among neonates delivered at the University of Gondar comprehensive specialized hospital, Northwest Ethiopia. *Frontiers in Pediatrics*. 2022;10. doi: https://www.frontiersin.org/journals/pediatrics/articles/10.3389/fped.2022.899922/full
7. S AM, Senguttuvan A, K D, Raghupathy NS. A study of maternal factors influencing birth weight in newborn in a tertiary care hospital. *International Journal of Contemporary Pediatrics*. 2021;8(11):1810-1814. doi: https://www.ijpediatrics.com/index.php/ijcp/article/view/4544
8. Shaohua Y, Bin Z, Mei L, et al. Maternal risk factors and neonatal outcomes associated with low birth weight. *Frontiers in Genetics*. 2022;13. doi: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.1019321/full
9. Prudhivi S, Bhosgi R. Maternal factors influencing low birth weight babies. *International Journal of Contemporary Pediatrics*. Published online 2015:287-296. doi: https://www.ijpediatrics.com/index.php/ijcp/article/view/401
10. He Z, Bishwajit G, Yaya S, Cheng Z, Zou D, Zhou Y. Prevalence of low birth weight and its association with maternal body weight status in selected countries in Africa: a cross-sectional study. *BMJ Open*. 2018;8(8):e020410. doi: https://bmjopen.bmj.com/content/8/8/e020410
11. Sarika DrM, Vishwakarma DrR, Rao DrR. "Study of low birth weight babies and their association with maternal risk factors." *Pediatric Review: International Journal of Pediatric*

Research. 2020;7(7):379-387. doi: https://pediatrics.medresearch.in/index.php/ijpr/article/view/632

12. Siddartha Aradhya, Tegunimataka A, Øystein Kravdal, et al. Maternal age and the risk of low birthweight and pre-term delivery: a pan-Nordic comparison. 2022;52(1):156-164. doi: https://academic.oup.com/ije/article/52/1/156/6814403
13. World Health Organization. Low birth weight. World Health Organization. Published 2025. https://www.who.int/data/nutrition/nlis/info/low-birth-weight
14. Cleveland Clinic. Low Birth Weight. Cleveland Clinic. Published 2023. https://my.clevelandclinic.org/health/diseases/24980-low-birth-weight
15. Data Sets. www.openintro.org. https://www.openintro.org/data/

APPENDIX

[321]:
```
import pandas as pd
```

IMPORTING NCBIRTHS DATA AND ITS INFORMATION

[323]:
```
data = pd.read_csv("ncbirths.csv")
```

[325]:
```
data
```

[325]:

	fage	mage	mature	weeks	premie	visits	marital	gained	weight	lowbirthweight	gender	habit	whitemom
0	NaN	13	younger mom	39.0	full term	10.0	not married	38.0	7.63	not low	male	nonsmoker	not white
1	NaN	14	younger mom	42.0	full term	15.0	not married	20.0	7.88	not low	male	nonsmoker	not white
2	19.0	15	younger mom	37.0	full term	11.0	not married	38.0	6.63	not low	female	nonsmoker	white
3	21.0	15	younger mom	41.0	full term	6.0	not married	34.0	8.00	not low	male	nonsmoker	white
4	NaN	15	younger mom	39.0	full term	9.0	not married	27.0	6.38	not low	female	nonsmoker	not white
...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	47.0	42	mature mom	40.0	full term	10.0	married	26.0	8.44	not low	male	nonsmoker	not white
996	34.0	42	mature mom	38.0	full term	18.0	married	20.0	6.19	not low	female	nonsmoker	white
997	39.0	45	mature mom	40.0	full term	15.0	married	32.0	6.94	not low	female	nonsmoker	white
998	55.0	46	mature mom	31.0	premie	8.0	married	25.0	4.56	low	female	nonsmoker	not white
999	45.0	50	mature mom	39.0	full term	14.0	married	23.0	7.13	not low	female	nonsmoker	white

1000 rows × 13 columns

[327]:
```
data.isnull().sum()
```

[327]:
```
fage              171
mage                0
mature              0
weeks               2
premie              2
visits              9
marital             1
gained             27
weight              0
lowbirthweight      0
gender              0
habit               1
whitemom            2
dtype: int64
```

DATA CLEANING

Since the fage column is considered to be a control variable and has nulls, it is being dropped

:
```
df1 = data.drop(columns=["fage"])
df1.head()
```

:

	mage	mature	weeks	premie	visits	marital	gained	weight	lowbirthweight	gender	habit	whitemom
0	13	younger mom	39.0	full term	10.0	not married	38.0	7.63	not low	male	nonsmoker	not white
1	14	younger mom	42.0	full term	15.0	not married	20.0	7.88	not low	male	nonsmoker	not white
2	15	younger mom	37.0	full term	11.0	not married	38.0	6.63	not low	female	nonsmoker	white
3	15	younger mom	41.0	full term	6.0	not married	34.0	8.00	not low	male	nonsmoker	white
4	15	younger mom	39.0	full term	9.0	not married	27.0	6.38	not low	female	nonsmoker	not white

Removing null values in weeks,visits,gained columns and imputing missing values using median()

:
```
df1['weeks'].fillna(df1['weeks'].median(), inplace=True)
df1['visits'].fillna(df1['visits'].median(), inplace=True)
df1['gained'].fillna(df1['gained'].median(), inplace=True)
```

```
/var/folders/vq/7yx0z70n6bg1b1n2pf7_xj800000gn/T/ipykernel_51211/92347378.py:1: FutureWarning: A value is trying to be
```

DATA CLEANING

Since the fage column is considered to be a control variable and has nulls, it is being dropped

```
df1 = data.drop(columns=["fage"])
df1.head()
```

	mage	mature	weeks	premie	visits	marital	gained	weight	lowbirthweight	gender	habit	whitemom
0	13	younger mom	39.0	full term	10.0	not married	38.0	7.63	not low	male	nonsmoker	not white
1	14	younger mom	42.0	full term	15.0	not married	20.0	7.88	not low	male	nonsmoker	not white
2	15	younger mom	37.0	full term	11.0	not married	38.0	6.63	not low	female	nonsmoker	white
3	15	younger mom	41.0	full term	6.0	not married	34.0	8.00	not low	male	nonsmoker	white
4	15	younger mom	39.0	full term	9.0	not married	27.0	6.38	not low	female	nonsmoker	not white

Removing null values in weeks,visits,gained columns and imputing missing values using median()

```
df1['weeks'].fillna(df1['weeks'].median(), inplace=True)
df1['visits'].fillna(df1['visits'].median(), inplace=True)
df1['gained'].fillna(df1['gained'].median(), inplace=True)
```

Removing null values in premie,marital,habit,whitemom columns and imputing missing values using mode()

```
df1['premie'].fillna(df1['premie'].mode()[0], inplace=True)
df1['marital'].fillna(df1['marital'].mode()[0], inplace=True)
df1['habit'].fillna(df1['habit'].mode()[0], inplace=True)
df1['whitemom'].fillna(df1['whitemom'].mode()[0], inplace=True)
```

[335]:
```
df1.isnull().sum()
```

[335]:
```
mage              0
mature            0
weeks             0
premie            0
visits            0
marital           0
gained            0
weight            0
lowbirthweight    0
gender            0
habit             0
whitemom          0
dtype: int64
```

[337]:
```
df1.nunique()
```

[337]:
```
mage               33
mature              2
weeks              23
premie              2
visits             26
marital             2
gained             71
weight            126
lowbirthweight      2
gender              2
habit               2
whitemom            2
dtype: int64
```

[339]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   mage            1000 non-null   int64
 1   mature          1000 non-null   object
 2   weeks           1000 non-null   float64
 3   premie          1000 non-null   object
 4   visits          1000 non-null   float64
 5   marital         1000 non-null   object
 6   gained          1000 non-null   float64
 7   weight          1000 non-null   float64
 8   lowbirthweight  1000 non-null   object
 9   gender          1000 non-null   object
 10  habit           1000 non-null   object
 11  whitemom        1000 non-null   object
dtypes: float64(4), int64(1), object(7)
memory usage: 93.9+ KB
```

df1 is a cleaned dataset and has 1000 rows x 12 columns

RENAMING COLUMNS

```
df1.rename(columns={
    'mage': 'Mother_Age',
    'mature': 'Mother_Maturity',
    'weeks': 'Gestation_Weeks',
    'premie': 'Birth_Term',
    'visits': 'Prenatal_Visits',
    'marital': 'Marital_Status',
    'gained': 'Mother_Weight_Gained',
    'weight': 'Baby_Weight',
    'lowbirthweight': 'Baby_Weight_Category',
    'gender': 'Baby_Gender',
    'habit': 'Smoking_Status'
}, inplace=True)
```

```
df = df1
df
```

	Mother_Age	Mother_Maturity	Gestation_Weeks	Birth_Term	Prenatal_Visits	Marital_Status	Mother_Weight_Gained	Baby_Weight	Baby_Weight_Category
0	13	younger mom	39.0	full term	10.0	not married	38.0	7.63	not low
1	14	younger mom	42.0	full term	15.0	not married	20.0	7.88	not low
2	15	younger mom	37.0	full term	11.0	not married	38.0	6.63	not low
3	15	younger mom	41.0	full term	6.0	not married	34.0	8.00	not low
4	15	younger mom	39.0	full term	9.0	not married	27.0	6.38	not low
...	...	...	...	...	...	...	...	...	...
995	42	mature mom	40.0	full term	10.0	married	26.0	8.44	not low
996	42	mature mom	38.0	full term	18.0	married	20.0	6.19	not low
997	45	mature mom	40.0	full term	15.0	married	32.0	6.94	not low
998	46	mature mom	31.0	premie	8.0	married	25.0	4.56	low
999	50	mature mom	39.0	full term	14.0	married	23.0	7.13	not low

1000 rows × 12 columns

df is a cleaned dataset with columns renamed for better understanding

---

CALCULATED FIELD

```
df['Weight_Gain_per_Week'] = df['Mother_Weight_Gained'] / df['Gestation_Weeks']
df['Weight_Gain_per_Week']
```

```
0      0.974359
1      0.476190
2      1.027027
3      0.829268
4      0.692308
         ...
995    0.650000
996    0.526316
997    0.800000
998    0.806452
999    0.589744
Name: Weight_Gain_per_Week, Length: 1000, dtype: float64
```

```
df
```

	Mother_Age	Mother_Maturity	Gestation_Weeks	Birth_Term	Prenatal_Visits	Marital_Status	Mother_Weight_Gained	Baby_Weight	Baby_Weight_Category
0	13	younger mom	39.0	full term	10.0	not married	38.0	7.63	not low
1	14	younger mom	42.0	full term	15.0	not married	20.0	7.88	not low
2	15	younger mom	37.0	full term	11.0	not married	38.0	6.63	not low
3	15	younger mom	41.0	full term	6.0	not married	34.0	8.00	not low
4	15	younger mom	39.0	full term	9.0	not married	27.0	6.38	not low
...	...	...	...	...	...	...	...	...	...
995	42	mature mom	40.0	full term	10.0	married	26.0	8.44	not low
996	42	mature mom	38.0	full term	18.0	married	20.0	6.19	not low
997	45	mature mom	40.0	full term	15.0	married	32.0	6.94	not low
998	46	mature mom	31.0	premie	8.0	married	25.0	4.56	low
999	50	mature mom	39.0	full term	14.0	married	23.0	7.13	not low

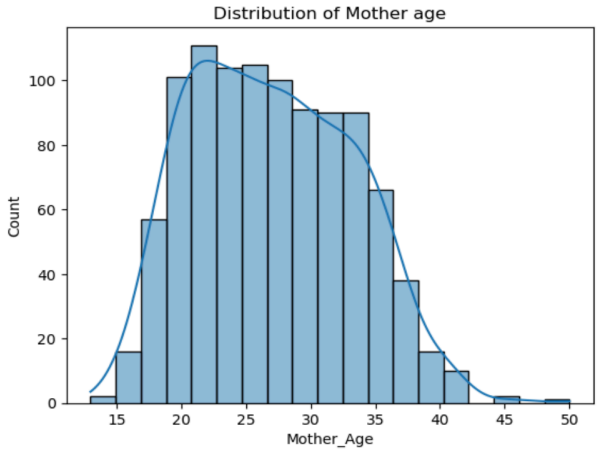
1000 rows × 13 columns

ENCODING VARIABLES AND DATA VISUALIZATION

[349]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

1. Histogram

[351]:
```
sns.histplot(df['Mother_Age'], kde = 'True')
plt.title('Distribution of Mother age')
plt.show()
```



Interpretation: Mother's age follows a right-skewed distribution, with most mothers aged between 20 and 35 years. There are fewer younger (<18) and older (>40) mothers in the dataset, which may influence how maternal age affects outcomes like birth weight.

---

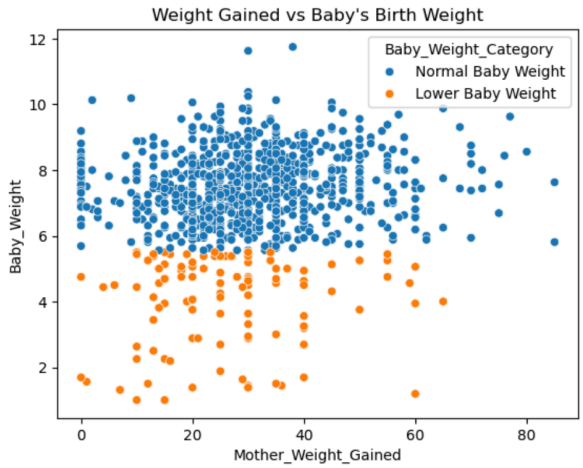
2. Scatter Plot

[355]:
```
df['Baby_Weight_Category'] = df['Baby_Weight_Category'].replace({
    'low': 'Lower Baby Weight',
    'not low': 'Normal Baby Weight'
})
```

[357]:
```
df['Baby_Weight_Category'].value_counts()
```

[357]:
```
Baby_Weight_Category
Normal Baby Weight    889
Lower Baby Weight     111
Name: count, dtype: int64
```

[359]:
```
sns.scatterplot(x='Mother_Weight_Gained', y='Baby_Weight', hue='Baby_Weight_Category', data=df)
plt.title("Weight Gained vs Baby's Birth Weight")
plt.show()
```

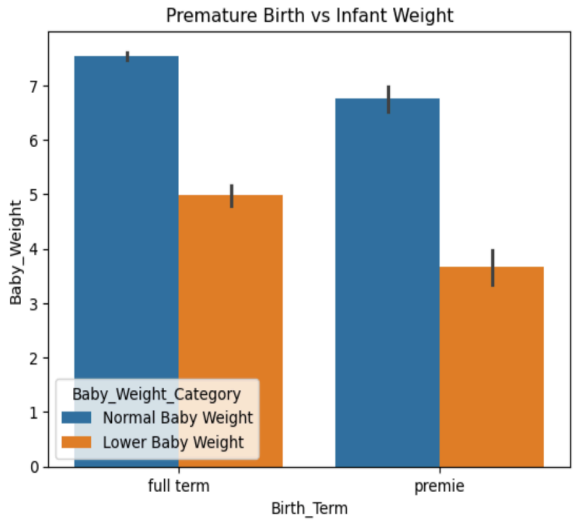


Interpretation: There's a visible separation between lower and normal baby weight categories. Mothers who gained more weight during pregnancy tend to have babies in the normal weight range.

---

3. Barplot

[361]:
```
sns.barplot(x='Birth_Term', y='Baby_Weight', hue='Baby_Weight_Category', data=df)
plt.title("Premature Birth vs Infant Weight")
plt.show()
```

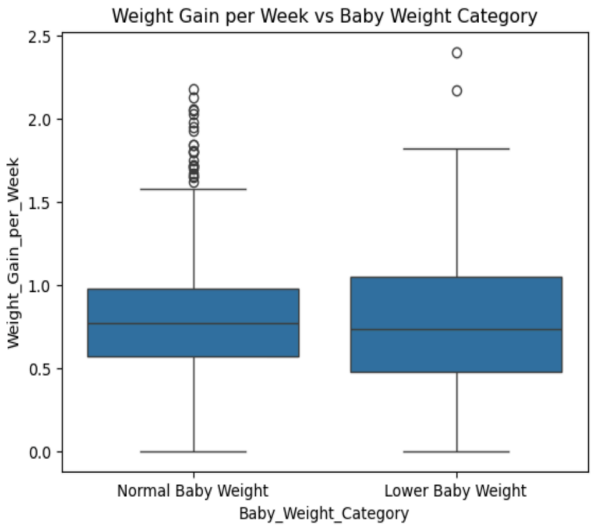


Interpretation: Infants born prematurely (premie) have a visibly lower average birth weight, especially among those classified as "Lower Baby Weight." This supports the well-established finding that premature delivery is a major risk factor for low birth weight.

---

4. Boxplot

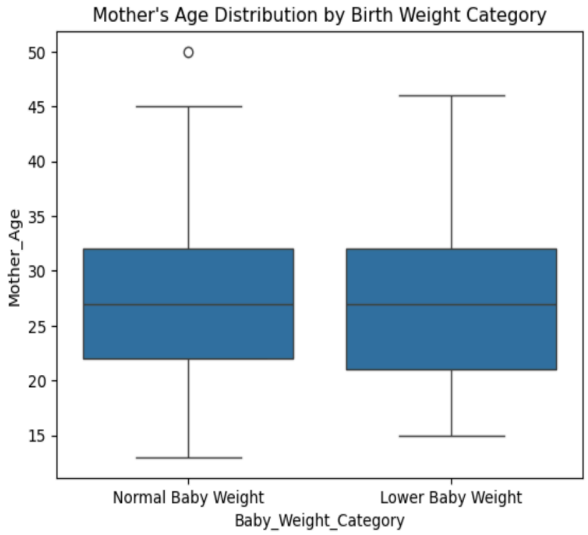
[363]:
```
sns.boxplot(x='Baby_Weight_Category', y='Weight_Gain_per_Week', data=df)
plt.title("Weight Gain per Week vs Baby Weight Category")
plt.show()
```



Interpretation: The boxplot shows that mothers with normal-weight babies generally gained slightly more weight per week than those with lower-weight babies. The median is higher for the normal group, and more high outliers are present.

---

[365]:
```
sns.boxplot(x='Baby_Weight_Category', y='Mother_Age', data=df)
plt.title("Mother's Age Distribution by Birth Weight Category")
plt.show()
```

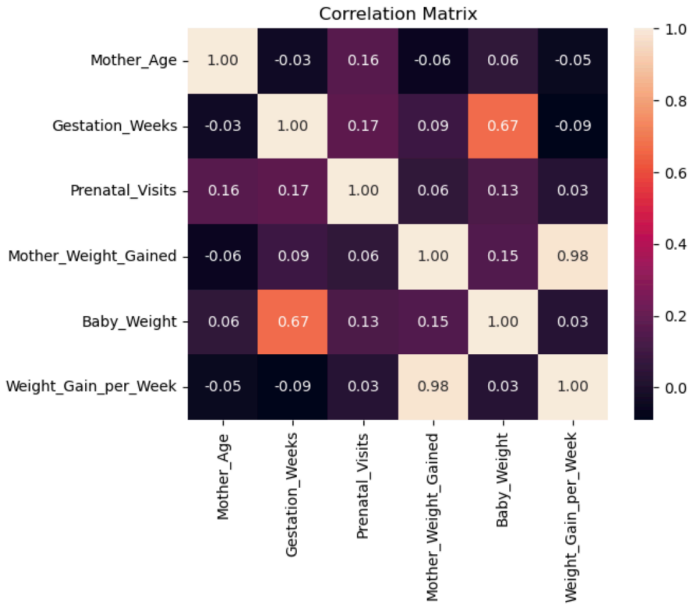


Interpretation: The distribution of maternal age appears similar across both baby weight categories. Median age and spread are nearly identical, suggesting no strong association between maternal age and low birth weight in this sample.

---

5. Heatmap

[367]:
```
sns.heatmap(df.select_dtypes(include='number').corr(), annot=True, fmt='.2f')
plt.title("Correlation Matrix")
plt.show()
```



Interpretation: The correlation matrix shows a moderate positive correlation (0.67) between gestational weeks and baby weight, suggesting that longer pregnancies are associated with higher birth weights. Other variables like prenatal visits and maternal weight gain show weak correlations with baby weight, indicating they may have a smaller individual effect.

Interpretation:

-The correlation heatmap shows the linear relationships among the variables in the dataset. Gestation Weeks has a strong positive correlation with Baby Weight (r = 0.66), suggesting that longer pregnancies are associated with heavier babies. It also shows a moderately strong negative correlation with Baby Weight Category (r = -0.58), indicating that shorter gestational periods are linked to a higher risk of low birth weight

-Baby Weight is highly negatively correlated with the Baby Weight Category (r = -0.72), which is expected, since lower weight leads to a higher chance of being categorized as Low Birth Weight.

-Mother's Weight Gained and Weight Gain per Week are very strongly positively correlated (r = 0.98), suggesting multicollinearity, one could potentially be dropped. Both have weak to moderate positive correlations with Baby Weight (r ≈ 0.15–0.16) and weak negative correlations with Baby Weight Category (r ≈ -0.11), hinting that more maternal weight gain may contribute to better birth weight outcomes.

-Prenatal Visits show a weak positive correlation with Baby Weight (r = 0.13) and a weak negative correlation with Baby Weight Category (r = -0.11), implying a mild protective role of prenatal care.

-Mother's Age has negligible correlations with both Baby Weight (r = 0.06) and Baby Weight Category (r = -0.01), indicating it may not be a strong standalone predictor in this dataset.

5]:
```
x = df[["Mother_Age", "Gestation_Weeks", "Prenatal_Visits", "Mother_Weight_Gained", "Mother_Maturity_younger mom", "Birth_Term_premie", "Marit
y = df["Baby_Weight_Category"]
```

7]:
```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from numpy import set_printoptions
test =SelectKBest(score_func=chi2, k=4)
fit=test.fit(x,y)
set_printoptions(precision=2)
print(fit.scores_)
```

```
[6.01e-03 7.69e+01 1.68e+01 6.57e+01 1.22e-01 2.66e+02 8.65e+00 1.30e+00
 1.86e+00]
```

9]:
```
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
rfe = RFE(model, n_features_to_select=1)
fit=rfe.fit(x,y)
print(fit.ranking_)
```

```
[8 1 7 9 5 3 2 4 6]
```

---

[401]:
```
from sklearn.ensemble import ExtraTreesClassifier
model = ExtraTreesClassifier()
model.fit(x,y)
print(model.feature_importances_)
```

```
[0.14 0.24 0.15 0.16 0.01 0.24 0.02 0.02 0.02]
```

Interpretation:

-Based on the results from three feature selection methods—Chi-Square, Recursive Feature Elimination (RFE), and ExtraTreesClassifier—the most important predictors for determining baby weight category are Birth_Term_premie, Gestation_Weeks, Mother_Weight_Gained, Prenatal_Visits, and Marital_Status_not married.

-Among these, Birth_Term_premie emerged as the most influential feature across all methods, indicating that premature births are strongly associated with lower baby weight.

-Gestation_Weeks also consistently ranked high, highlighting the importance of pregnancy duration in determining newborn weight. Mother_Weight_Gained and Prenatal_Visits reflect maternal nutrition and healthcare access, both of which are critical factors in fetal development.

-Additionally, Marital_Status_not married may act as a proxy for social and economic factors that indirectly influence birth outcomes. These five features not only showed strong statistical relevance but also align with medical knowledge, making them well-suited for building a predictive model for baby weight classificatio

MODEL EVALUATION

[403]:
```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report, confusion_matrix, roc_auc_score, roc_curve
from sklearn.metrics import accuracy_score
```

[405]:
```
X = df[["Birth_Term_premie", "Gestation_Weeks", "Mother_Weight_Gained", "Prenatal_Visits", "Marital_Status_not married"]]
# Target Variable
y = df["Baby_Weight_Category"]
```

[407]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
```

---

1. LOGISTIC REGRESSION

[439]:
```
from sklearn.linear_model import LogisticRegression
```

[441]:
```
#Fit the model
log_model = LogisticRegression()
log_model.fit(X_train, y_train)
```

[441]:
```
▾ LogisticRegression
LogisticRegression()
```

[443]:
```
# Predict and evaluate
log_y_pred = log_model.predict(X_test)
conf_matrix = confusion_matrix(y_test, log_y_pred)
class_report = classification_report(y_test, log_y_pred)
```

[445]:
```
print("Confusion Matrix:\n", conf_matrix)
print("\nClassification Report:\n", class_report)
```

```
Confusion Matrix:
 [[177   1]
 [ 10  12]]

Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.99      0.97       178
           1       0.92      0.55      0.69        22

    accuracy                           0.94       200
   macro avg       0.93      0.77      0.83       200
weighted avg       0.94      0.94      0.94       200
```

[447]:
```
from sklearn.metrics import roc_auc_score
log_y_proba = log_model.predict_proba(X_test)[:, 1]
log_auc = roc_auc_score(y_test, log_y_proba)
print(f"Logistic Regression AUC-ROC: {log_auc:.2f}")
```

```
Logistic Regression AUC-ROC: 0.89
```

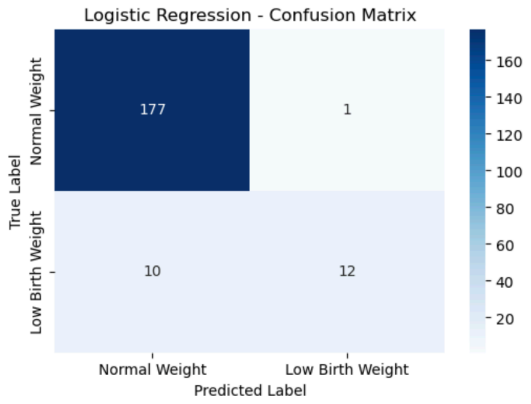
[449]:
```
# Accuracy
logistic_accuracy = accuracy_score(y_test, log_y_pred)
logistic_accuracy_percent = logistic_accuracy * 100
print(f"Logistic Regression Accuracy: {logistic_accuracy_percent:.2f}%")
```

```
Logistic Regression Accuracy: 94.50%
```

Toggle output scrolling

Interpretation: Logistic Regression offered the highest accuracy among all models and serves as a strong baseline. It helped interpret the influence of maternal features in a linear framework. However, its moderate recall indicates that while it predicts most normal weight babies correctly, it could miss some low-birth-weight cases.

[451]:
```
cm = confusion_matrix(y_test, log_y_pred)
plt.figure(figsize=(6, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
            xticklabels=['Normal Weight', 'Low Birth Weight'],
            yticklabels=['Normal Weight', 'Low Birth Weight'])
plt.title('Logistic Regression - Confusion Matrix')
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.show()
```



---

2. RANDOM FOREST

[453]:
```
from sklearn.ensemble import RandomForestClassifier
```

[455]:
```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

[457]:
```
# Fit the model
rf_model = RandomForestClassifier(n_estimators=100, class_weight='balanced')
rf_model.fit(X_train_scaled, y_train)
```

[457]:
```
RandomForestClassifier(class_weight='balanced')
```

[459]:
```
#Predict and evaluate
y_pred = rf_model.predict(X_test_scaled)
y_proba = rf_model.predict_proba(X_test_scaled)[:, 1]
```

[461]:
```
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))

print("\nClassification Report:")
print(classification_report(y_test, y_pred))
```

```
Confusion Matrix:
[[176   2]
 [ 10  12]]

Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.99      0.97       178
           1       0.86      0.55      0.67        22

    accuracy                           0.94       200
   macro avg       0.90      0.77      0.82       200
weighted avg       0.94      0.94      0.93       200
```

[463]:
```
rf_y_proba = rf_model.predict_proba(X_test_scaled)[:, 1]
rf_auc = roc_auc_score(y_test, rf_y_proba)
print(f"Random Forest AUC-ROC: {rf_auc:.2f}")
```

```
Random Forest AUC-ROC: 0.88
```

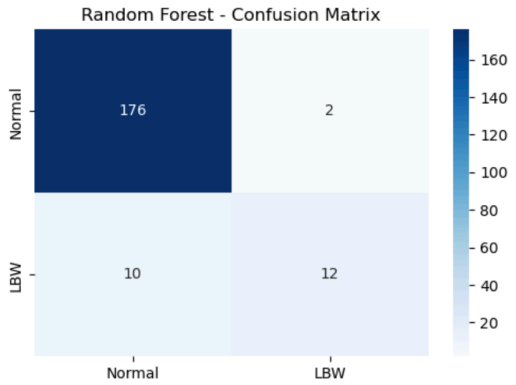
[465]:
```
rf_accuracy = accuracy_score(y_test, y_pred)
rf_accuracy_percent = rf_accuracy * 100
print(f"Random Forest Accuracy: {rf_accuracy_percent:.2f}%")
```

```
Random Forest Accuracy: 94.00%
```

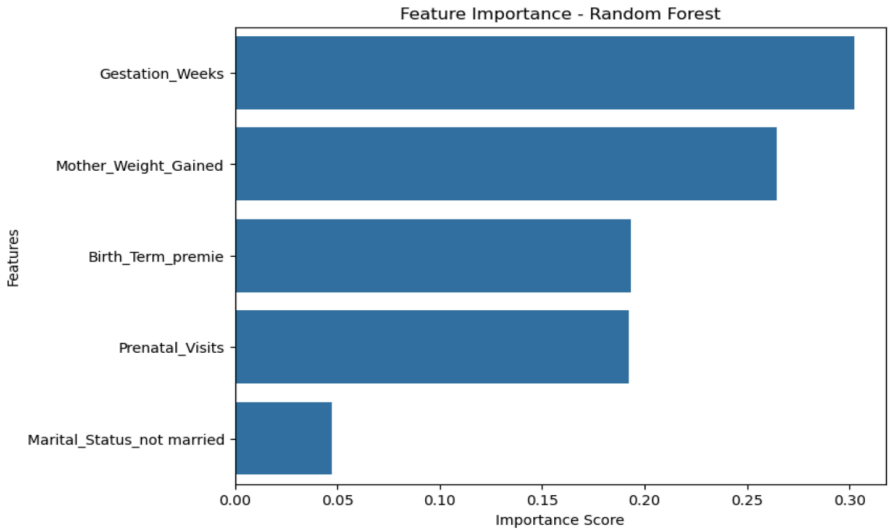
Interpretation: Random Forest effectively captured non-linear relationships and feature interactions. It achieved the best balance between precision and recall, making it a highly dependable model, especially in medical settings where both sensitivity and specificity are important. Feature importance analysis further enhanced interpretability.

[467]:
```
cm = confusion_matrix(y_test, y_pred)

# Plot the heatmap
plt.figure(figsize=(6, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['Normal', 'LBW'], yticklabels=['Normal', 'LBW'])
plt.title('Random Forest - Confusion Matrix')
plt.show()
```



[469]:
```
feature_importances = pd.Series(rf_model.feature_importances_, index=X.columns).sort_values(ascending=False)
plt.figure(figsize=(8,6))
sns.barplot(x=feature_importances.values, y=feature_importances.index)
plt.title("Feature Importance - Random Forest")
plt.xlabel("Importance Score")
plt.ylabel("Features")
plt.show()
```



3. SVM

[471]:
```
from sklearn.svm import SVC
```

[473]:
```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

[475]:
```
#Fit the model
svm_model = SVC()
svm_model.fit(X_train_scaled, y_train)
```

[475]:
```
SVC()
```

[477]:
```
print("\nConfusion Matrix:")
print(confusion_matrix(y_test, svm_y_pred))

print("\nClassification Report:")
print(classification_report(y_test, svm_y_pred))
```

```
Confusion Matrix:
[[177   1]
 [ 11  11]]

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.99      0.97       178
           1       0.92      0.50      0.65        22

    accuracy                           0.94       200
   macro avg       0.93      0.75      0.81       200
weighted avg       0.94      0.94      0.93       200
```

[479]:
```
svm_y_scores = svm_model.decision_function(X_test_scaled)
svm_auc = roc_auc_score(y_test, svm_y_scores)
print(f"SVM AUC-ROC: {svm_auc:.2f}")
```

```
SVM AUC-ROC: 0.90
```

[483]:
```
#Predict and evaluate
svm_y_pred = svm_model.predict(X_test_scaled)
svm_accuracy = accuracy_score(y_test, svm_y_pred)
print(f"SVM Accuracy: {svm_accuracy * 100:.2f}%")
```

```
SVM Accuracy: 94.00%
```

XGboost

[487]:
```
import xgboost as xgb
```

[489]:
```
xgb_model = xgb.XGBClassifier(use_label_encoder=False, eval_metric='logloss')
```

[491]:
```
xgb_model.fit(X_train, y_train)
```

```
/opt/anaconda3/lib/python3.12/site-packages/xgboost/training.py:183: UserWarning: [17:23:39] WARNING: /Users/runner/work/xgboost/xgboost/src/
learner.cc:738:
Parameters: { "use_label_encoder" } are not used.

  bst.update(dtrain, iteration=i, fobj=obj)
```

[491]:
```
XGBClassifier
                colsample_bylevel=None, colsample_bynode=None,
                colsample_bytree=None, device=None, early_stopping_rounds=None,
                enable_categorical=False, eval_metric='logloss',
                feature_types=None, feature_weights=None, gamma=None,
                grow_policy=None, importance_type=None,
                interaction_constraints=None, learning_rate=None, max_bin=None,
                max_cat_threshold=None, max_cat_to_onehot=None,
                max_delta_step=None, max_depth=None, max_leaves=None,
                min_child_weight=None, missing=nan, monotone_constraints=None,
                multi_strategy=None, n_estimators=None, n_jobs=None,
                num_parallel_tree=None, ...)
```

[493]:
```
xgb_y_pred = xgb_model.predict(X_test)
```

[495]:
```
print("\nConfusion Matrix:")
print(confusion_matrix(y_test, xgb_y_pred))

print("\nClassification Report:")
print(classification_report(y_test, xgb_y_pred))
```

```
Confusion Matrix:
[[174   4]
 [ 11  11]]

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.98      0.96       178
           1       0.73      0.50      0.59        22

    accuracy                           0.93       200
   macro avg       0.84      0.74      0.78       200
weighted avg       0.92      0.93      0.92       200
```

7]:
```
xgb_y_proba = xgb_model.predict_proba(X_test)[:, 1]
xgb_auc = roc_auc_score(y_test, xgb_y_proba)
print(f"XGBoost AUC-ROC: {xgb_auc:.2f}")
```

```
XGBoost AUC-ROC: 0.88
```

9]:
```
xgb_accuracy = accuracy_score(y_test, xgb_y_pred)
print(f"XGBoost Accuracy: {xgb_accuracy * 100:.2f}%")
```

```
XGBoost Accuracy: 92.50%
```

Interpretation: XGBoost delivered competitive performance with efficient training and robust handling of feature interactions. While it achieved a good precision, its slightly lower recall compared to Random Forest limits its sensitivity toward detecting low birth weight cases.

---

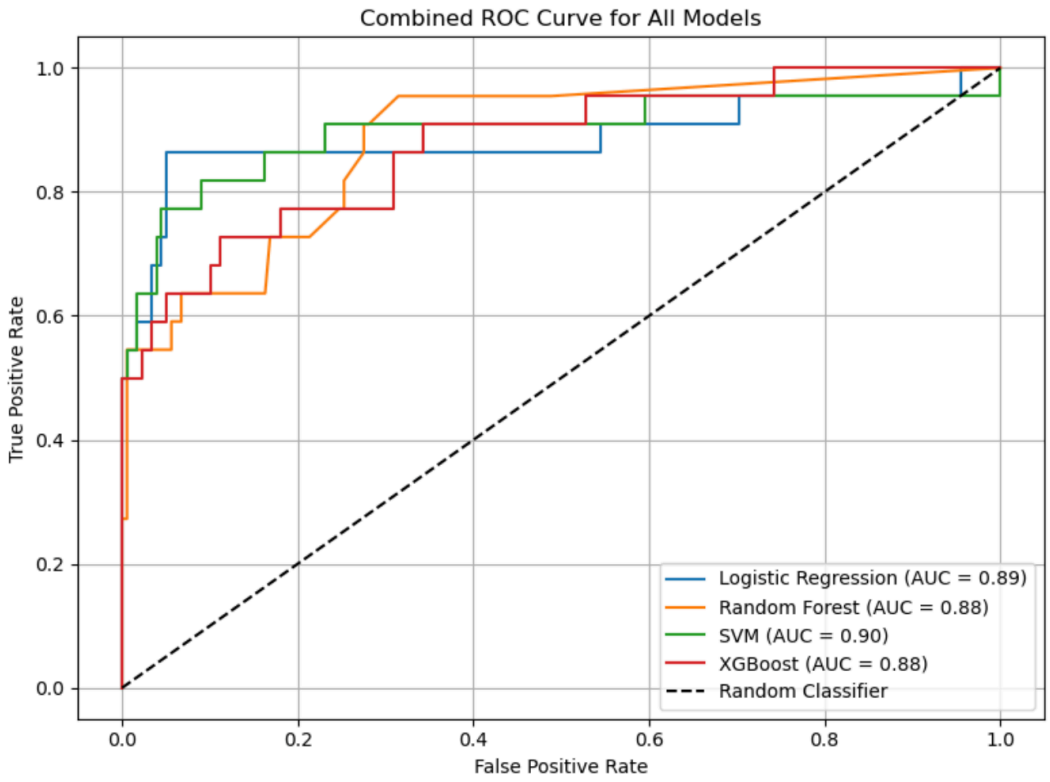
[501]:
```
# Logistic Regression
fpr_log, tpr_log, _ = roc_curve(y_test, log_y_proba)

# Random Forest
fpr_rf, tpr_rf, _ = roc_curve(y_test, rf_y_proba)

# SVM
fpr_svm, tpr_svm, _ = roc_curve(y_test, svm_y_scores)

# XGBoost
fpr_xgb, tpr_xgb, _ = roc_curve(y_test, xgb_y_proba)

plt.figure(figsize=(8, 6))
plt.plot(fpr_log, tpr_log, label=f'Logistic Regression (AUC = {log_auc:.2f})')
plt.plot(fpr_rf, tpr_rf, label=f'Random Forest (AUC = {rf_auc:.2f})')
plt.plot(fpr_svm, tpr_svm, label=f'SVM (AUC = {svm_auc:.2f})')
plt.plot(fpr_xgb, tpr_xgb, label=f'XGBoost (AUC = {xgb_auc:.2f})')
plt.plot([0, 1], [0, 1], 'k--', label='Random Classifier')
plt.title('Combined ROC Curve for All Models')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend(loc='lower right')
plt.grid(True)
plt.tight_layout()
plt.show()
```



LOADING DATA

Hide

```
NCbirths.2004 <- read.csv(file = "/Users/chandinikotha/Desktop/Capstone")
```

```
Warning in file(file, "rt") :
  'raw = FALSE' but '/Users/chandinikotha/Desktop/Capstone' is not a regular file
Warning in file(file, "rt") :
  cannot open file '/Users/chandinikotha/Desktop/Capstone': it is a directory
Error in file(file, "rt") : cannot open the connection
```

Hide

```
summary(object = NCbirths.2004)
```

```
      mage          mature              weeks          premie              visits       marital              gained
 Min.   :13   Length:1000        Min.   :20.00   Length:1000        Min.   : 0.0   Length:1000        Min.   : 0.
00
 1st Qu.:22   Class :character   1st Qu.:37.00   Class :character   1st Qu.:10.0   Class :character   1st Qu.:21.
00
 Median :27   Mode  :character   Median :39.00   Mode  :character   Median :12.0   Mode  :character   Median :30.
00
 Mean   :27                      Mean   :38.34                      Mean   :12.1                      Mean   :30.
32
 3rd Qu.:32                      3rd Qu.:40.00                      3rd Qu.:15.0                      3rd Qu.:38.
00
 Max.   :50                      Max.   :45.00                      Max.   :30.0                      Max.   :85.
00
     weight       lowbirthweight        gender             habit             whitemom
 Min.   : 1.000   Length:1000        Length:1000        Length:1000        Length:1000
 1st Qu.: 6.380   Class :character   Class :character   Class :character   Class :character
 Median : 7.310   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   : 7.101
 3rd Qu.: 8.060
 Max.   :11.750
```

```
Hmisc::describe(x = NCbirths.2004)
```

```
NCbirths.2004

 12  Variables      1000  Observations
--------------------------------------------------------------------------------------------------------------------
-------------------
mage
       n  missing distinct     Info     Mean  pMedian      Gmd      .05      .10      .25      .50      .75
.90      .95
    1000        0       33    0.998       27       27    7.103       18       19       22       27       32
35       37

lowest : 13 14 15 16 17, highest: 41 42 45 46 50
--------------------------------------------------------------------------------------------------------------------
-------------------
mature
       n  missing distinct
    1000        0        2

Value       mature mom younger mom
Frequency           133          867
Proportion        0.133        0.867
--------------------------------------------------------------------------------------------------------------------
-------------------
weeks
       n  missing distinct     Info     Mean  pMedian      Gmd      .05      .10      .25      .50      .75
.90      .95
    1000        0       23    0.974    38.34     38.5    2.846       33       35       37       39       40
41       42

lowest : 20 22 24 25 26, highest: 41 42 43 44 45
--------------------------------------------------------------------------------------------------------------------
-------------------
premie
       n  missing distinct
    1000        0        2

Value      full term    premie
Frequency          848        152
Proportion       0.848        0.152
```

```
premie
       n  missing distinct
    1000        0        2

Value      full term    premie
Frequency         848       152
Proportion      0.848     0.152
-------------------------------------------------------------------------------------------------------------------
-------------------
visits
       n  missing distinct     Info     Mean  pMedian      Gmd      .05      .10      .25      .50      .75
.90      .95
    1000        0       26    0.989     12.1       12    4.246        5        7       10       12       15
16       18

lowest :  0  2  3  4  5, highest: 23 24 25 26 30
-------------------------------------------------------------------------------------------------------------------
-------------------
marital
       n  missing distinct
    1000        0        2

Value          married not married
Frequency          614          386
Proportion       0.614        0.386
-------------------------------------------------------------------------------------------------------------------
-------------------
gained
       n  missing distinct     Info     Mean  pMedian      Gmd      .05      .10      .25      .50      .75
.90      .95
    1000        0       71    0.997    30.32       30    15.45       10       14       21       30       38
50       56

lowest :  0  1  2  3  4, highest: 75 76 77 80 85
-------------------------------------------------------------------------------------------------------------------
-------------------
weight
       n  missing distinct     Info     Mean  pMedian      Gmd      .05      .10      .25      .50      .75
.90      .95
    1000        0      126        1    7.101     7.22    1.595     4.44     5.44     6.38     7.31     8.06
8.75     9.13
```

```
lowest : 1     1.19  1.31  1.38  1.44 , highest: 10.19 10.25 10.38 11.63 11.75
-------------------------------------------------------------------------------------------------------------------
-------------------
lowbirthweight
       n  missing distinct
    1000        0        2

Value          low not low
Frequency      111     889
Proportion   0.111   0.889
-------------------------------------------------------------------------------------------------------------------
-------------------
gender
       n  missing distinct
    1000        0        2

Value      female   male
Frequency     503    497
Proportion  0.503  0.497
-------------------------------------------------------------------------------------------------------------------
-------------------
habit
       n  missing distinct
    1000        0        2

Value      nonsmoker    smoker
Frequency        874       126
Proportion     0.874     0.126
-------------------------------------------------------------------------------------------------------------------
-------------------
whitemom
       n  missing distinct
    1000        0        2

Value      not white     white
Frequency        284       716
Proportion     0.284     0.716
-------------------------------------------------------------------------------------------------------------------
-------------------
```

#####STATISTICAL TESTS##########

INDEPENDENT t test

1. Null Hypothesis: There is no difference in mean baby weight between full-term and premature births
2. Alternative Hypothesis: There is a difference in mean baby weight between full-term and premature births
3. Statistical test:

premie <chr>	mean_weight <dbl>	var_weight <dbl>	n <int>
full term	7.454575	1.170198	848
premie	5.128421	3.879265	152

2 rows

Hide

```
print(t_test_result)
```

```
    Welch Two Sample t-test

data:  weight by premie
t = 14.182, df = 167.68, p-value < 2.2e-16
alternative hypothesis: true difference in means between group full term and group premie is not equal to 0
95 percent confidence interval:
 2.002351 2.649958
sample estimates:
mean in group full term    mean in group premie
               7.454575                5.128421
```

4. INTERPRETATION: The Welch's two-sample t-test resulted a t-statistic of 14.18 with approximately 167.68 degrees of freedom, and a p-value < 2.2e-16, indicating that the observed difference in birth weights between full-term and premature babies is highly statistically significant. The 95% confidence interval for the difference in means is [2.00, 2.65], which does not include zero, reinforcing the statistical evidence against the null hypothesis. Babies born full-term have a substantially higher mean birth weight (M = 7.45 lbs) compared to those born prematurely (M = 5.13 lbs), with less variability in weight among the full-term group. We reject the null hypothesis and conclude that birth term is significantly associated with infant birth weight, with full-term delivery linked to higher birth weights in this sample.

CHI SQUARE TEST

Between premie(BIRTH TERM) - weight(BABY BIRTH WEIGHT):

1.Null Hypothesis: There is no association between birth term (premature/full-term) and baby weight category (low/normal). 2.Alternative Hypothesis: There is a significant association between birth term and baby weight category. 3. Statistical test:

Hide

```
table(NCbirths.2004$premie, NCbirths.2004$lowbirthweight)
```

Hide

```
chisq.test(table(NCbirths.2004$premie, NCbirths.2004$lowbirthweight))
```

4. INTERPRETATION: The chi-square test revealed a test statistic of $\chi^2 = 308.37$ with 1 degree of freedom and a p-value less than 2.2e-16. This extremely small p-value indicates that the probability of observing such a strong association between premature birth status and low birth weight, assuming the null hypothesis is true. Since the p-value < 0.05, we reject the null hypothesis and conclude that there is statistically significant evidence of a relationship between birth term and low birth weight. Babies born prematurely are much more likely to have low birth weight compared to those born full term.

Between habit(SMOKING STATUS) - weight(BABY BIRTH WEIGHT):

1. Null Hypothesis: There is no association between maternal smoking status and baby weight category.
2. Alternative Hypothesis: There is an association between maternal smoking status and baby weight category.
3. Statistical test:

Hide

```
table(NCbirths.2004$habit, NCbirths.2004$lowbirthweight)
```

```
           low not low
  nonsmoker  93     781
  smoker     18     108
```

Hide

```
chisq.test(table(NCbirths.2004$habit, NCbirths.2004$lowbirthweight))
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  table(NCbirths.2004$habit, NCbirths.2004$lowbirthweight)
X-squared = 1.1363, df = 1, p-value = 0.2864
```

4. INTERPRETATION: The chi-square test examining the association between maternal smoking status and baby weight category yielded a test statistic of $\chi^2 = 1.14$ with 1 degree of freedom and a p-value of 0.29. Since the p-value > 0.05, we fail to reject the null hypothesis. This result suggests that there is no statistically significant association between whether a mother smoked during pregnancy and the likelihood of delivering a low birth weight baby.

ANOVA

Between mature(MOTHER MATURITY STATUS) - weight(BABY BIRTH WEIGHT):

1. Null Hypothesis: Mean baby weight is the same across maturity groups.
2. Alternate Hypothesis: Mean baby weight differs across maturity groups.
3. Statistical test:

Hide

```
summary(anova_result2)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
mature        1    0.1  0.0926   0.041   0.84
Residuals   998 2274.3  2.2788
```


```
             Df Sum Sq Mean Sq F value Pr(>F)
mature        1    0.1  0.0926   0.041   0.84
Residuals   998 2274.3  2.2788
```


4. INTERPRETATION: The one-way ANOVA was conducted to compare the effect of maternal maturity status (mature) on baby weight (weight). The test returned F(1, 998) = 0.041, with a p-value = 0.84. Since the p-value > 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant difference in mean baby weight between mature and younger mothers in this sample.

Between gained(MOTHER WEIGHT GAINED) - premie(BIRTH TERM):

1. Null Hypothesis: Mean weight gain is equal across birth term (premie) groups.
2. Alternate Hypothesis: Mean weight gain is different across birth term (premie) groups.
3. Statistical test:

Hide

```
summary(anova_result3)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
premie        1   3351    3351   17.26 3.55e-05 ***
Residuals   998 193788     194
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4. INTERPRETATION: The ANOVA test shows a significant difference in maternal weight gain between full-term and premature births (F(1, 998) = 17.26, p < 0.001). Mothers of full-term babies gained more weight on average than those with premature deliveries. This suggests gestational length may influence weight gain.