A Deep Dive: Exploring the Impact of Maternal Characteristics on Preterm Birth through Multiple and Logistic Regression Analysis

EXECUTIVE SUMMARY

I. Problem & Hypothesis

- Research Question: What maternal characteristics (e.g. age, ethnicity, education, socioeconomic status) significantly influence the likelihood of preterm birth?

- Data analysis through multiple regression and logistic regression modeling will allow for an objective assessment of the relationship between maternal characteristics and preterm birth.

- There are no definitive reasons for preterm births. However, there are risk factors that can increase the likelihood of women having a preterm birth.

- Some factors for delivering a preterm baby in the past include being pregnant with multiple fetuses, tobacco and/or substance abuse (*Premature Birth*, 2022).

- By identifying specific maternal factors related to preterm birth, healthcare providers can implement targeted interventions, monitoring, and care plans to mitigate the risk and improve patient outcomes.

- **Null hypothesis**: There is no significant association between individual maternal characteristics and the likelihood of preterm birth.

- **Alternate Hypothesis**: There is a significant relationship between individual maternal characteristics and the likelihood of preterm birth.

II. Data Analysis Process

- Data collection: Data obtained from NCHS (National Center for Health Statistics) database based on birth certificate regression. Data sourced from an existing database on Kaggle.

- Data extraction: relevant variables related to maternal characteristics and preterm birth were selected from the natality database.
    - Independent variables: birth_year, birth_month, birth_time, birth_place, father_age, mother_education, marital_status, mother_age, father_education , interval_llb , cigarettes, mother_height, mother_bmi, pre_preg_weight, pre_preg_diabetes, gest_diabetes, pre_preg_hypertension, gest_hypertension, infertility_treatment, prev_cesarian, gonorrhea, syphilis, chlamydia, hepatitis_b , hepatitis_c, labor_induction, labor_augmentation, steroids, antibiotics, chorioamnionitis, anesthesia, apgar5, apgar10, plurality, gender, infant_weight
    - Dependent variable: prev_preterm_birth

- Data Cleaning: the data was collected by examining and cleaning data to remove duplicates, errors, or missing variables.

- Data preparation: one-hot encoding, univariate analysis, bivariate analysis, heat maps

- Data analysis:
    - Normality test completed:
        - Kolmogorov-Smirnov Test
            - KS statistic: 0.10665330380822602
            - p-value: 0.19116286378085373

        - Shapiro-Wilk Test
            - Shapiro-Wilk test statistic: 0.17307919263839722
            - P-value: 0.0

    - Multiple regression analysis: used to assess the relationship between individual maternal characteristics and preterm birth
        - utilized p-values >0.05 and VIF to reduce to the initial model to the final regression model
        - standardized coefficients
        - found coefficients and intercept
        - created residual plot
        - cross-validation scores
        - cook's distance
        - ANOVA

    - Logistic Regression analysis: coefficients indicated the direction and magnitude of association between the maternal characteristics and preterm birth
        - AIC of the initial and reduced model
        - cross-validation of the logistic regression model
        - confusion matrix
        - found coefficients and intercept
        - classification report
        - ROC Curve
        - variable importance (permutation importance method)
        -
- Model Comparison
    - comparison of logistic regression model and random forest metrics
    - Regulation technique

III. Findings

- The following maternal characteristics associated with preterm birth that were identified to be statistically significant were birthplace, mother age, mother education, father education, the interval between last live birth, cigarettes use, delivery weight, use of steroids, plurality, infant weight, marital status, pre-pregnancy diabetes, gestational diabetes, pre-pregnancy hypertension, gestational hypertension, infertility treatment, gonorrhea, hepatitis C, labor induction, labor augmentation, antibiotics usage, and gender.

- The multiple regression analysis indicated an overall good model at predicting preterm birth.
    - MSE= 0.313 suggesting the predicted values deviated from the actual values by a small amount
    - RSE= 0.177 suggesting there is some remaining amount of variation in the data that is not explained by the multiple regression model

- The null hypothesis of this analysis was rejected.

- The logistic regression analysis revelated the following independent variable to be positive: steroid use, marital status, diabetes prior to pregnancy, hypertension prior to pregnancy, gestational hypertension, infertility treatments, gonorrhea, hepatitis C, labor augmentation, antibiotic use, and gender.
    - The negative coefficients were birthplace, mother's education, father's education, the interval between the last live birth, cigarette use, delivery weight, plurality, infant weight, and labor induction.

- The logistic regression model struggled to correctly identify instances of preterm birth indicated (positive class) by a low recall, precision, and F1-score. However, the model performed well in predicting term births (negative class)

- The logistic model does not identify any instances of preterm birth, only term births. The estimated probability of preterm birth for this given set of predictor values is 0.56%.

- Variable importance demonstrated the independent variables had limited impact on predicting preterm birth.

- Selected maternal characteristics might not be strong predictors of preterm birth and/or the model needs to be improved to better capture the relationship between maternal characteristics and preterm birth.

IV. Limitations

- The sourced data from Kaggle used in this analysis had limited variables compared to the original NCHS dataset.

- The dataset used only contains data for one year which limits the ability to capture trends and generalize findings over time.

- multicollinearity is a limitation in multiple regression analysis. This can occur when independent variables are highly correlated with each other making it difficult to determine the independent effects on the dependent variable.

- The analysis focused on finding correlations between maternal characteristics and preterm birth, but it is difficult to establish causation due to potential confounding factors.

- There is limited information on maternal and paternal characteristics and the absence of socioeconomic factors that could contribute to preterm birth.

- The regression models used in the analysis rely on assumptions such as linearity, independence, and absence of multicollinearity. violations of these assumptions can affect the validity of the results of the data.

- The limited number of independent variables used in the analysis may not fully capture all the factors that influence preterm birth.

- This analysis does not account for the variability in underlying factors contributing to preterm birth among women who have had a previous preterm birth.

V. Recommended Actions

- The models did well at predicting term births but not preterm births in relation to maternal characteristics.

- It is recommended that model be expanded with more independent variables.
  - Other possible factors influencing preterm birth that could include race, substance abuse, inadequate prenatal care, healthcare resource availability, or hormonal imbalances.

- Further investigation should be conducted to determine maternal characteristics influencing preterm birth to understand specific factors and characteristics that contribute to preterm birth to provide valuable insights on prevention and intervention strategies.

- It is recommended there be external validation. Validating the findings and model on another dataset that is similar should be done to assess the generalizability of the results.

- One future approach for further studying would be to complete a longitudinal analysis.
  - A longitudinal analysis can provide insight into patterns and changes in risk factors associated with preterm birth and offer multiple time points during pregnancy and during

postpartum care for a dynamic assessment and investigation into maternal characteristics.

- Another approach for future study would be to look further into causal inferences. Using casual inference methods can establish a causal relationship between specific risk factors and preterm birth.

VI. Benefits of Study

- Through multiple regression and logistic regression modeling the analysis allowed for an objective evaluation of the relationship between maternal characteristics and preterm birth.

- The study identified serval maternal characteristics that were found to be significant predictors of preterm birth including birthplace, mother's age, mother's education, father's education, tobacco use, delivery weight, use of steroids, and various medical conditions. These findings contribute to the understanding of the factors associated with preterm birth.

- By identifying specific maternal factors related to preterm birth, healthcare providers can implement targeted interventions, monitoring, and care plan to mitigate risk and improve patient outcomes. This can help reduce the incidence of preterm birth and improve the health and patient outcomes of both the mother and baby.

- This study rejects the null hypothesis that there is no significant association between individual maternal characteristics and the likelihood of preterm birth.
    - The findings support the alternative hypothesis and provide evidence for the existence of significant relationships between maternal characteristics and preterm birth, prompting further investigation.

- This analysis emphasizes the complex nature of predicting preterm birth and highlights the need for further research and model improvement.
    - By recognizing the limitations of this analysis, researchers can refine their methods and explore additional variables that may influence preterm birth such as race, prenatal care, healthcare resource availability, or hormonal imbalances.

- This analysis contributes to the existing body of knowledge on preterm birth by providing insights into the favors influencing critical health outcomes.

Panopto Video Presentation Link:
https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=85b2b8b5-dd13-4b4e-853a-b036011ff1f3