Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Investigating the Impact of Maternal Characteristics and Smoking on Birth Weight: An Ensemble Regression Analysis

K DeviPriya, Anusuri Krishna Veni, Manjula Devarakonda Venkata, G Muni Nagamani, Velchuri Balaji, N V Ramya Devi Kotla

Department of Computer Science and Engineering, Lakireddy Bali Reddy College of Enigneering, JNTUK, Andhrapradesh, India, k.devipriya20@gmail.com
[2]Department of Computer Science & Engineering Data Science, Madanapalle Institute of Technology & Science
[3]Department of CSE, Pragati Engineering College (A), Surampalem, Affliated to JNTUK, AP, India
[4]Department of Computer Science &Engineering, Andhra Loyola Institute of Engineering and Technology, Vijayawada, Andhra Pradesh-520008
[5]K L E F Deemed to Be University, Green Fields, Vaddeswaram, Guntur(dt) Andhra Pradesh, 522302
[6]Dept of IT SRKR Engineering College, Bhimavaram, Andhrapradesh, India

Low birth weight is a significant public health concern associated with increased risk of infant mortality and long-term health issues. Understanding the maternal factors that influence birth weight is crucial for developing effective interventions to improve neonatal outcomes. This study investigates the relationships between birth weights and maternal characteristics including gestation period, age, height, weight, and smoking status. The analysis begins with data preprocessing, including handling missing values through imputation techniques. Then descriptive statistics and correlation analysis are computed to summarize the central tendencies, variability and correlation between features in the dataset. An ensemble regression model is employed to assess the influence of gestation period, age, height, weight, and smoking status on birth weight. An ensemble Random Forest regressor model performed high R2 score. The presented work in this study provides valuable insights into the factors influencing birth weight, emphasizing the significant impact of maternal smoking and predicting birth weight.

Keywords: Birth Weight, Maternal Health, Socioeconomic Status, Smoking during Pregnancy, Regression algorithms, Ensemble Regressors

1. Introduction

Low birth weight (LBW), defined as a weight of less than 2,500 grams [1] at birth, is a critical public health concern globally, contributing to increased infant mortality, morbidity, and long-term developmental issues. LBW is influenced [2-3] by various maternal factors, including age, nutritional status, smoking habits, and the length of gestation. Identifying at-risk pregnancies early on is crucial for implementing timely interventions to improve birth outcomes. However, traditional methods of predicting LBW based on clinical observations often lack accuracy and fail to account for the complex interplay of contributing factors. In recent years, the application of machine learning techniques [4-5] has emerged as a powerful tool for enhancing predictive models in healthcare. By leveraging large datasets and sophisticated algorithms, machine learning can uncover intricate patterns and relationships among various predictors that may not be apparent through conventional analysis [6]. In the context of predicting low birth weight, machine learning models can integrate diverse maternal characteristics, such as age, height, weight, and smoking status, to improve the accuracy of predictions and enable more personalized healthcare interventions [7]. This study explores the use of machine learning algorithms to predict low birth weight based on maternal characteristics with a focus on improving the accuracy and reliability of predictions. We investigated multiple regression models [8] including Linear Regression [9], Support Vector Regression (SVR)[10], Random Forest[11], Gradient Boosting[12], and using different data imputation methods included mean imputation, k-nearest neighbors (KNN) imputation, and Multiple Imputation by Chained Equations (MICE). We evaluated the performance of each regressor across the different imputation methods [13] using the R-squared metric, which measures the proportion of variance explained by the model. The remaining sections of paper is organized as related work in section 2,proposed methodology in section 3,experimental setup and results in section 4,and conclusion in section 5.

2. Related Work

Cho, Hannah, et al.[14] discussed key predictors of adverse birth outcomes in very low birth weight (VLBW) infants, with a particular focus on particulate matter concentration (PM10). The research utilized data from 10,423 VLBW infants drawn from the Korean Neonatal Network database, covering the period from January 2013 to December 2017. Five specific adverse birth outcomes were assessed as dependent variables: gestational age under 28 weeks, gestational age under 26 weeks, birth weight below 1000 grams, birth weight below 750 grams, and small-for-gestational age.Tessema, Zemenu Tadesse, et al.[15] focused on evaluating the prevalence and determinants of low birth weight across Sub-Saharan countries. Here, utilized the Kids Record (KR) dataset, which includes data on children under five years old born within the five years prior to the survey, specifically in the selected enumeration areas, and who had recorded birth weight data. To identify the factors influencing low birth weight, a multivariable mixed-effects logistic regression model was employed. Ghaderighahfarokhi et al. [16] analyzed secondary data from 450 medical records of newborns at educational hospitals associated with Ilam University of Medical Sciences. The birth records reviewed covered the period from April 2015 to April 2016. Data collection was conducted using a checklist that included two sections: demographic information and influential factors, encompassing 13

medical and neonatal factors, 4 maternal lifestyle factors, and 8 additional maternal factors. The data were analyzed using SPSS version 21 and WEKA software.Eliyati, Ning, et al.[17] uses Support Vector Machines (SVMs), a widely recognized algorithm in machine learning, to classify low birth weight (LBW) data. The primary goals of this study are to predict LBW classifications in Indonesia using SVMs and to compare the performance of SVMs with binary logistic regression, which is traditionally the most common model used for LBW data classification. Tao, Jing, et al. [18] integrated multiple electronic medical records with B-ultrasonic examinations of pregnant women to develop a hybrid classifier for predicting birth weight using Long Short-Term Memory (LSTM) networks. The clinical dataset includes information from 5,759 Chinese women who have given birth, encompassing over 57,000 obstetric electronic medical records. Emmanuel, Tlamelo, et al. discussed[19] an imputation experiment was done on the KNN and RF algorithms for imputation on the Iris and novel ID fan datasets to demonstrate how popular imputation algorithms perform. Samad, Manar D. et al. introduced novel approaches to enhance the Multiple Imputation by Chained Equations (MICE) algorithm, a widely successful method for missing value imputation (MVI), by substituting its linear regressions with non-linear regressions through the use of ensemble learning, deep learning, and clustering techniques[20-21].

3. Methodology

In this study, we focused on predicting low birth weight by leveraging a dataset containing maternal characteristics such as gestation period, parity, age, height, weight, and smoking status. The dataset consists of 1236 cases with variables such as birth weight, gestation period, parity, age, height, weight, and smoking status. Figure 1 depicted the overall architecture of the proposed work. The first step in our analysis involved addressing missing data. We applied three imputation methods: mean imputation, K-nearest neighbors (KNN) imputation, and Multiple Imputation by Chained Equations (MICE). Data imputation is a crucial step in handling missing data, which is common in healthcare datasets. Mean imputation replaces missing values with the average of observed values, providing a simple yet effective approach. KNN imputation considers the nearest neighbors to estimate missing values, while MICE iteratively imputes missing data using predictions from other variables, offering a more sophisticated method. These methods ensured that the dataset was complete for model training. After handling missing values, we normalized the continuous variables, which is essential for regression models to function optimally. The dataset was then split into an 80-20 train-test configuration, where the larger portion was used to train the models and the smaller portion was held back for evaluating their performance. We trained four regression models Random Forest, Gradient Boosting, Linear Regression, and Support Vector Regression (SVR) and optimized their hyper-parameters using grid search coupled with cross-validation. This approach ensured that the models were fine-tuned to provide the best possible performance on the test set. The evaluation of these models was based on the R-squared metric, which measures the proportion of variance in the target variable explained by the model. Initially identified missing values in the data then preprocessed using three types of imputation techniques of each feature then visualized before imputation and after imputation of each feature in Figure 2 and Figure 3. Then Statistical summaries provided essential insights into the dataset by offering a snapshot of central tendencies, spread, and distribution to understand

nature of the data. The count reveals the number of non-missing entries, helping assess data completeness. The mean indicates the average value, reflecting the central tendency, though it can be influenced by outliers. Standard deviation measures variability, with higher values indicating more spread in the data. The minimum and maximum values define the range, showing the extremes in the dataset, while the 25th and 75th percentiles (quartiles) reveal the distribution of the middle 50% of data, offering insights into skewness. The median provides a robust central value that isn't affected by outliers, and the mode shows the most frequently occurring value, highlighting common patterns mentioned in Table 1. Figure 4 shows histograms of birth weight based on Smoking Status and Figure5,Figure 6 visualizes correlation matrix that shows relationships between birth weight (bwt) and various maternal characteristics such as gestation period, parity, age, height, weight, and smoking status. From the generate values it was observer that moderate positive relationship between gestation and birth weight with a values 0.41,moderate negative relationship between smoke and birth weight with a value -0.21,height and weight features also positively correlated with birth weight with a value 0.20 and 0.15 ,case and parity features were negatively correlated with a value -0.06 and -0.05 shown in Table 2.

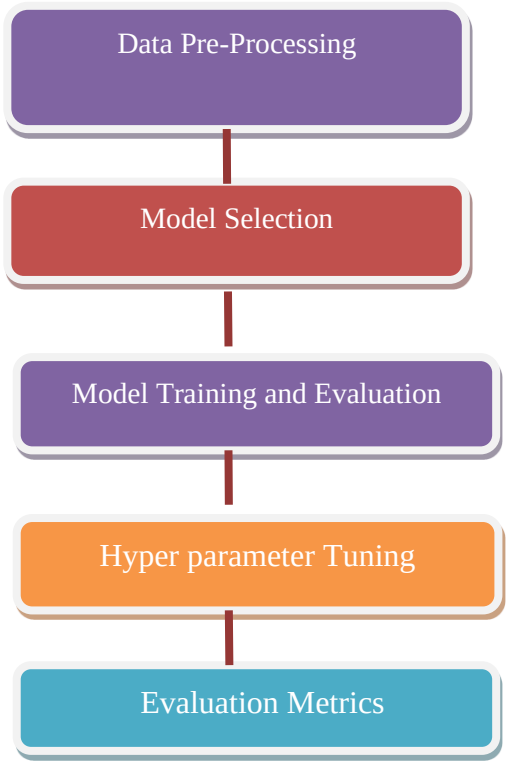

Figure1.Proposed Architecture

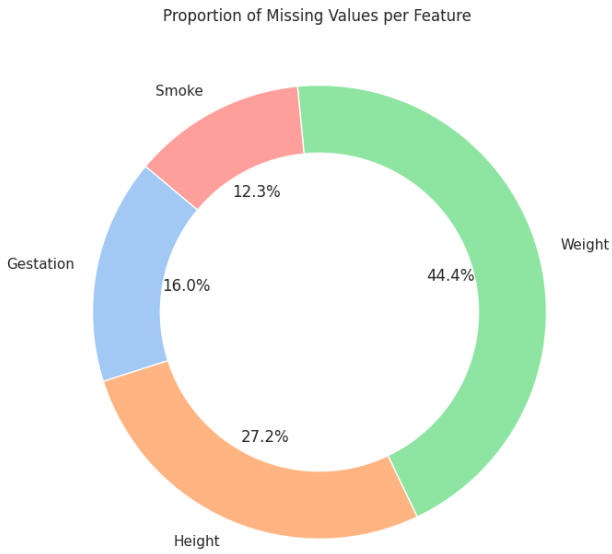

Figure 2.Percentage of missing values for each feature

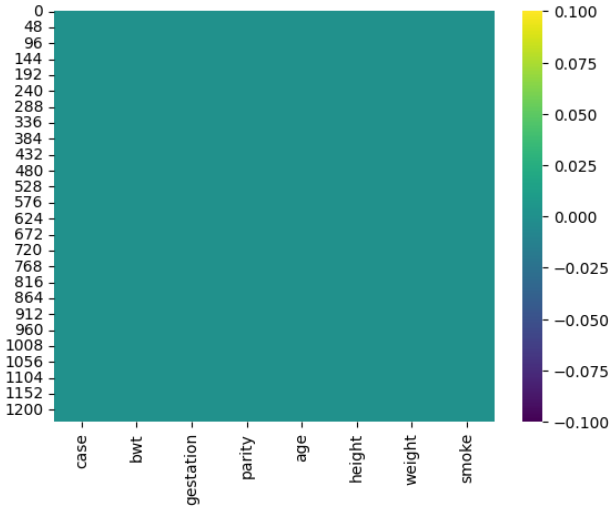

Figure 3.Replacement of Missing Values

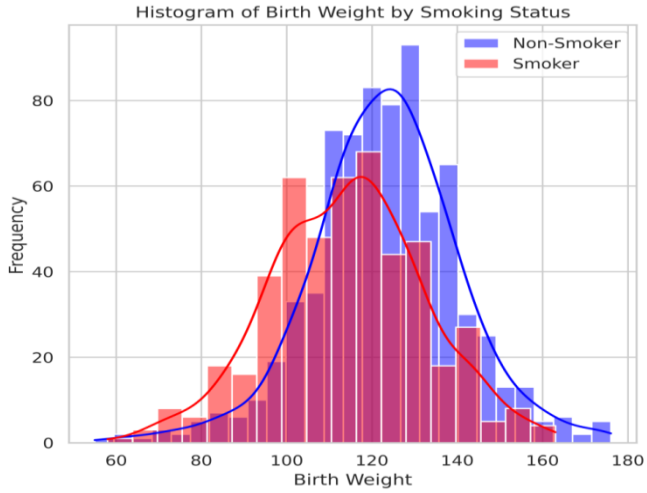

Figure 4.Histograms of Birth Weight Based on Smoking Status

Table 1.Statistics Description

	case	bwt	gestation	parity	age	height	weight	smoke
count	1236.0	1236.0	1236.0	1236.0	1236.0	1236.0	1236.0	case
mean	618.50	119.57	279.33	0.254	27.25	64.04	128.625833	0.394780
std	356.946775	18.236452	15.943114	0.435956	5.776722	2.510743	20.663939	0.487019
min	1.000000	55.000000	148.000000	0.000000	15.000000	53.000000	87.000000	0.000000
25%	309.750000	108.750000	272.00	0.000000	3.000000	62.000000	115.000000	0.000000
50%	618.500000	120.000000	280.000000	0.000000	26.000000	64.000000	126.000000	0.000000
75%	927.250000	131.000000	288.000000	1.000000	31.000000	66.000000	138.000000	1.000000
max	1236.000000	176.000000	353.000000	1.000000	45.000000	72.000000	250.000000	1.000000
mode	1.000000	115.000000	282.000000	0.000000	23.000000	64.000000	130.000000	0.000000

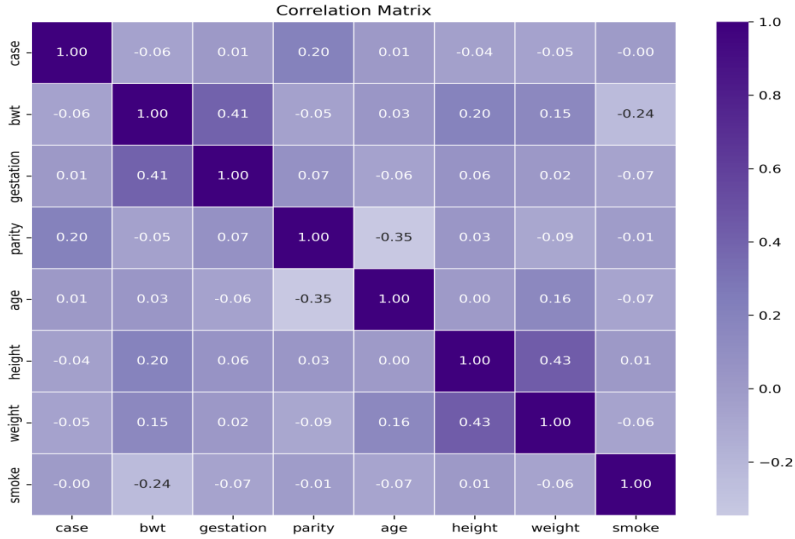

Figure 5. correlation matrix

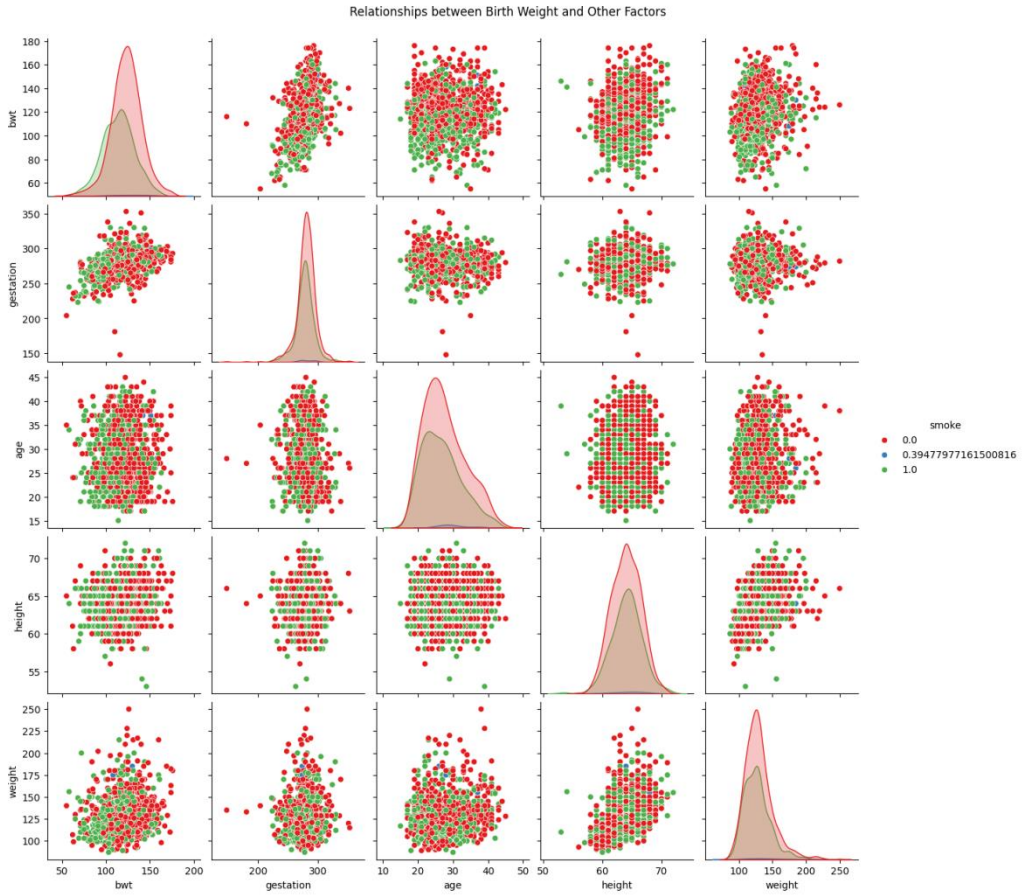

Figure 6. Relationship among birth weight and other factors

4. Experimental Setup and Results

In the experimental setup, we utilized Google Colab as the primary platform due to its cloud-based infrastructure, offering a convenient environment to execute Python code and access computational resources GPUs. This enabled faster training and optimization of machine learning models. Initially univariate analysis was performed to understand the impact individual feature on birth weight then regression models were trained on data and evaluated the Coefficient of Determination(R2).

Univariate Analysis

The analysis of average birth weight, beginning with a focus on maternal age, reveals that birth weight remains relatively consistent across various age groups. Most of the average birth weights fall between 115 and 125 grams. Younger mothers, especially those aged 18 to 22, tend to have slightly lower average birth weights, ranging from 107 to 120 grams, suggesting a minor association between younger maternal age and lower birth weights. For older mothers, particularly those over 35, average birth weights remain steady, ranging between 116 and 123

grams. This consistency suggests that advanced maternal age does not have a notable effect on birth weight, challenging concerns about older maternal age being linked to low birth weight shown in Figure 7.

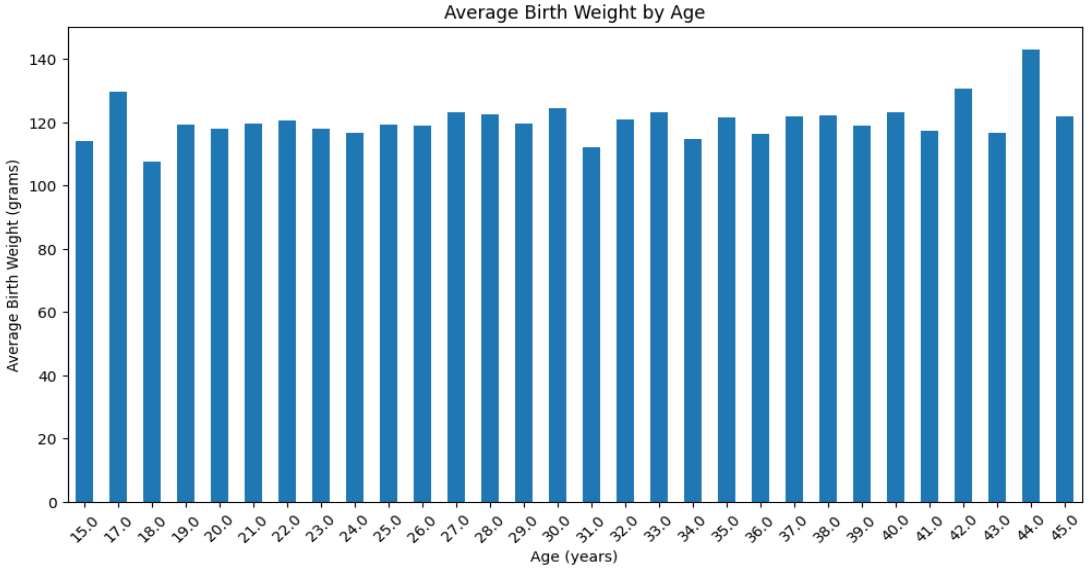

Figure 7. Average Birth Weight by Feature Age

Figure 8. indicated the average birth weight of babies categorized by the mother's height, ranging from 53 to 72 inches. As height increases, there is a general trend of rising birth weight, although some fluctuations are evident. At the lower end of the height spectrum (53 to 57 inches), the average birth weight fluctuates, with notably low values at heights like 56 inches (105g) and 57 inches (99g).From heights of 58 inches onwards, there is a more consistent increase, with slight variations. For instance, the average birth weight at 58 inches is 116.9g, rising steadily through the mid-60s, peaking around 70 inches at 131.46g.There is an overall upward trend, suggesting a positive correlation between maternal height and birth weight, though this isn't strictly linear as slight dips appear at heights like 69 and 72 inches. The analysis of average birth weight by smoking status reveals a noticeable difference between non-smokers and smokers. On average, the birth weight for babies born to non-smoking mothers is around 123 grams, while for smoking mothers, the average is lower at approximately 114 grams. This suggests that smoking during pregnancy may be associated with a reduction in birth weight. The observed difference of nearly 9 grams indicates that maternal smoking could have a negative impact on fetal growth, supporting concerns that smoking may lead to lower birth weights depicted in Figure 9. The analysis of average birth weight by parity shows that there is little to no significant difference between mothers who have given birth previously (parity 1) and first-time mothers (parity 0). The average birth weight for both groups is nearly equal, with only a slight variation. This indicates that parity does not have a substantial impact on birth weight, suggesting that whether a mother is giving birth for the first time or has had prior pregnancies do not significantly affect the baby's birth weight shown in Figure 10.

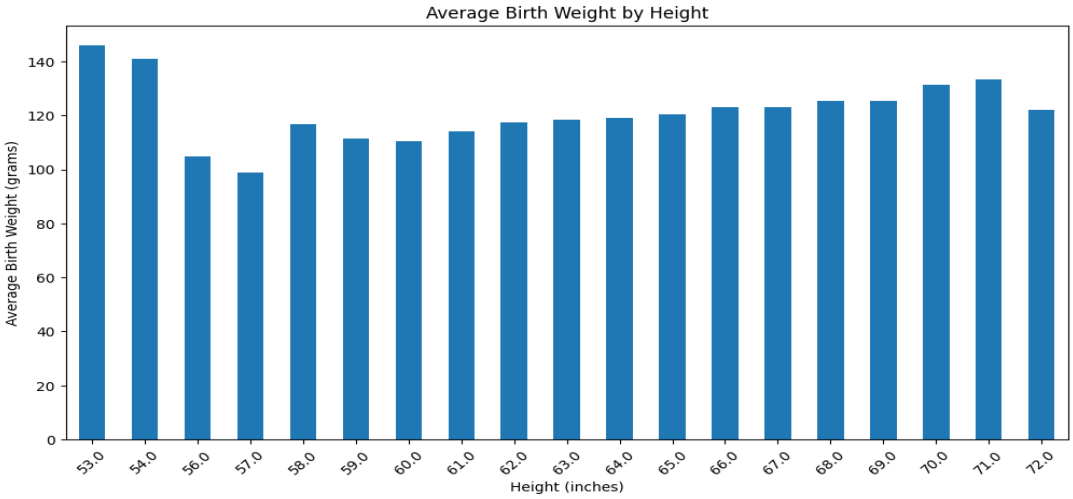

Figure 8. Average Birth Weight of Babies Categorized By the Mother's Height

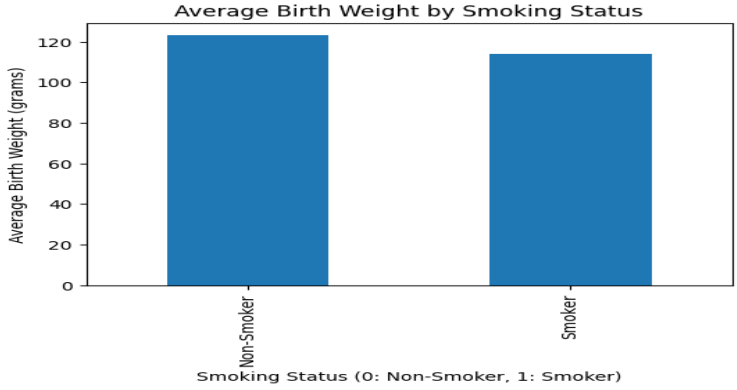

Figure 9. Average Birth Weight of Babies Categorized By the Mother's Smoking Status

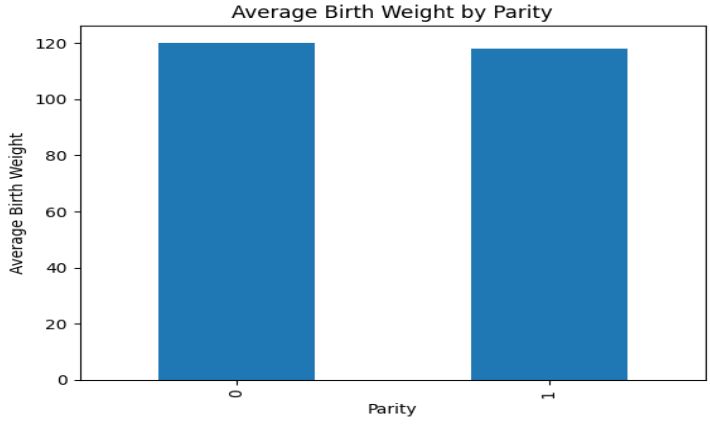

Figure 10. Average Birth Weight of Babies Categorized By the Mother's Smoking Status

R2 (Coefficient of Determination)

R2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It indicates how well the regression model fits the observed data. An R2 value closer to 1 suggests a better fit.R2 is measures as difference of 1- ration of sum of squared residuals (SSr) and total sum of squares(SSt)

$$R2 = 1 - SSr/SSt$$

Table 1,Figure 11. And Figure 12. represents each imputation method, regressor technique and R-squared values.Form the generated results higher R-squared value was generated by Random Forest regressor by integration of imputation method followed with MICE and KNN through Random Forest only.R-squared of 0.88, indicating its ability to capture complex relationships even when some values are replaced with the mean.The lower R-squared values observed for Gradient Boosting, Linear Regression, and SVR suggest that these models may not adequately handle the biases introduced by mean imputation, as this approach might overlook important patterns in datasets with inherent variability.

Table 1 R-squared Values of Different Regression Models Using Various Imputation Methods

Imputation Method	Regressor	R-squared
MEAN	RandomForest	0.8800
MEAN	GradientBoosting	0.3230
MEAN	LinearRegression	0.2453
MEAN	SVR	0.2600
KNN	RandomForest	0.6134
KNN	GradientBoosting	0.4003
KNN	LinearRegression	0.2459
KNN	SVR	0.2434
MICE	RandomForest	0.6114
MICE	GradientBoosting	0.4002
MICE	LinearRegression	0.2451
MICE	SVR	0.2422

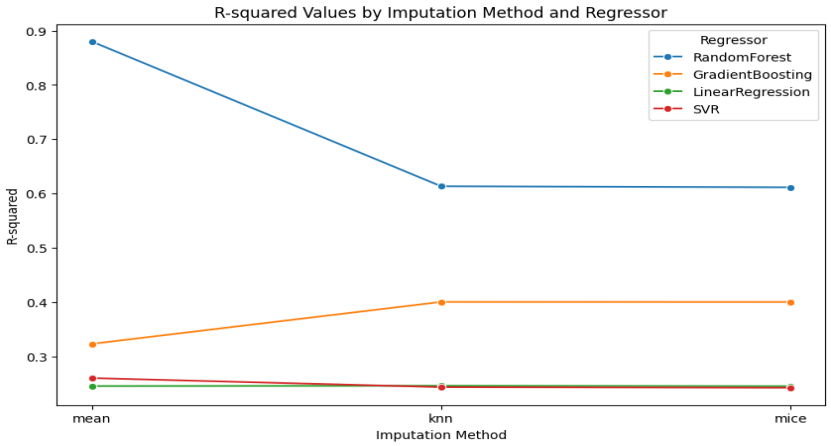

Figure 11. Comparison of Repressors

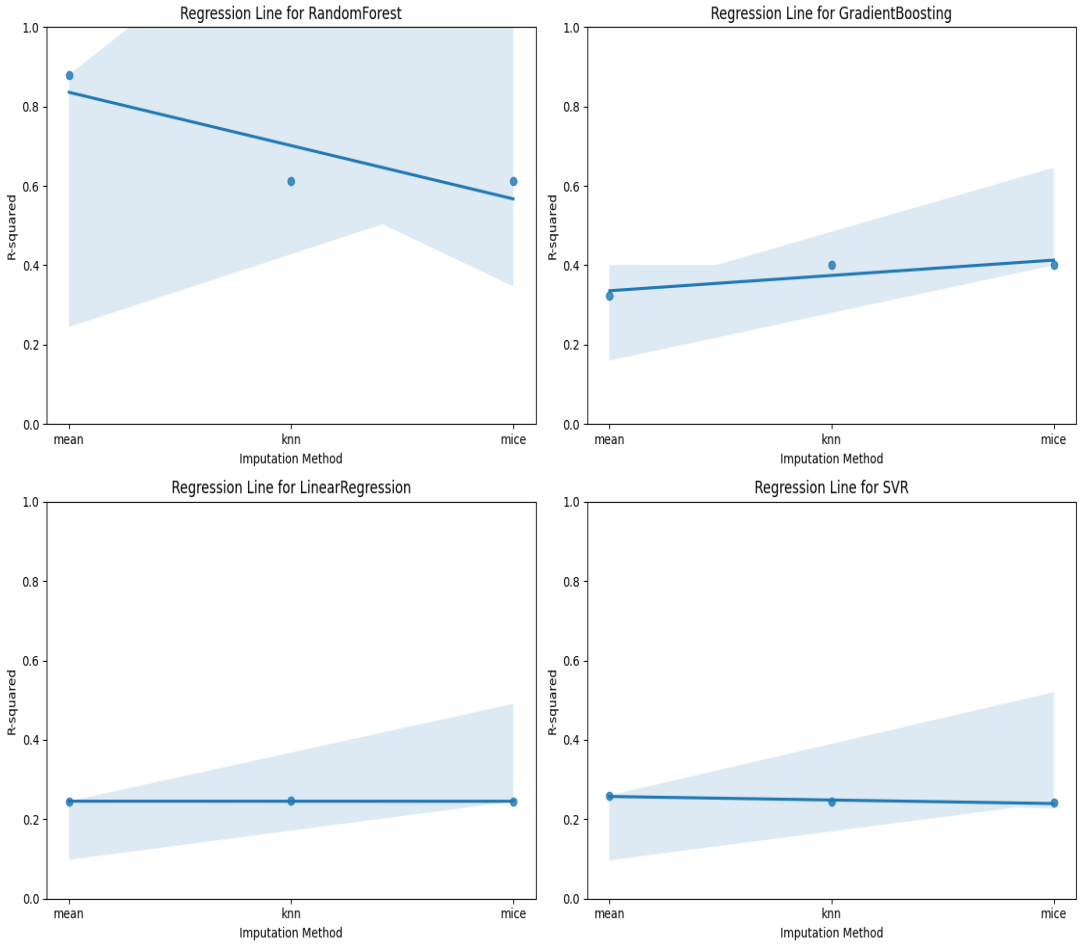

Figure 12. Regression lines of Models

5. Conclusion

In conclusion, this study demonstrates the effectiveness of a systematic approach to predicting low birth weight by incorporating essential steps such as data preprocessing, addressing missing values, conducting correlation analysis, and applying regression models. Imputation techniques Mean , KNN and MICE played a pivotal role in enhancing the accuracy of predictions, as they effectively handled missing data. Correlation analysis and Univaraite analysis helped identify the key features influencing birth weight, which informed the choice of models. The results show that Random Forest with mean imputation is the most effective approach, achieving the highest R-squared value of 0.88, indicating it captures the relationships in the data more effectively than other models and imputation strategies. Gradient Boosting, Linear Regression, and Support Vector Regression (SVR) exhibit lower predictive power, particularly when paired with mean imputation. These models generally show improved but still modest performance when using KNN and MICE imputation, though

none reach the level of Random Forest with mean imputation. These results emphasize the critical role of comprehensive data preparation and thoughtful model selection in improving predictive performance for low birth weight in newborns.

References

1. Singh, G., R. Chouhan, and K. Sidhu. "Maternal factors for low birth weight babies." Medical Journal Armed Forces India 65.1 (2009): 10-12.
2. Yadav, D. K., U. Chaudhary, and N. Shrestha. "Risk factors associated with low birth weight." (2011).
3. Hidalgo-Lopezosa, P., et al. "Sociodemographic factors associated with preterm birth and low birth weight: A cross-sectional study." Women and Birth 32.6 (2019): e538-e543.
4. Holzinger, Andreas. Machine learning for health informatics. Springer International Publishing, 2016.
5. Habehh, Hafsa, and Suril Gohel. "Machine learning in healthcare." Current genomics 22.4 (2021): 291.
6. Kourou, Konstantina, et al. "A machine learning-based pipeline for modeling medical, socio-demographic, lifestyle and self-reported psychological traits as predictors of mental health outcomes after breast cancer diagnosis: An initial effort to define resilience effects." Computers in Biology and Medicine 131 (2021): 104266.
7. Borson, Najmus Sakib, et al. "Correlation analysis of demographic factors on low birth weight and prediction modeling using machine learning techniques." 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). IEEE, 2020.
8. Maulud, Dastan, and Adnan M. Abdulazeez. "A review on linear regression comprehensive in machine learning." Journal of Applied Science and Technology Trends 1.2 (2020): 140-147.
9. Filzmoser, Peter, and Klaus Nordhausen. "Robust linear regression for high-dimensional data: An overview." Wiley Interdisciplinary Reviews: Computational Statistics 13.4 (2021): e1524.
10. Liu, Qinming, et al. "A new support vector regression model for equipment health diagnosis with small sample data missing and its application." Shock and Vibration 2021.1 (2021): 6675078.
11. Loef, Bette, et al. "Using random forest to identify longitudinal predictors of health in a 30-year cohort study." Scientific Reports 12.1 (2022): 10372.
12. Chumachenko, Dmytro, et al. "Investigation of statistical machine learning models for COVID-19 epidemic process simulation: Random forest, K-nearest neighbors, gradient boosting." Computation 10.6 (2022): 86.
13. Platias, Christos, and Georgios Petasis. "A comparison of machine learning methods for data imputation." 11th Hellenic Conference on Artificial Intelligence. 2020.
14. Morais, Flávio Leandro, et al. "Utilization of Tree-Based Machine Learning Models for Predicting Low Birth Weight Cases." (2024).
15. Tessema, Zemenu Tadesse, et al. "Prevalence of low birth weight and its associated factor at birth in Sub-Saharan Africa: A generalized linear mixed model." PloS one 16.3 (2021): e0248417.
16. Ghaderighahfarokhi, Shiva, Jamil Sadeghifar, and Mossayeb Mozafari. "A model to predict low birth weight infants and affecting factors using data mining techniques." Journal of Basic Research in Medical Sciences 5.3 (2018): 1-8.
17. Eliyati, Ning, et al. "Support vector machines for classification of low birth weight in Indonesia." Journal of Physics: Conference Series. Vol. 1282. No. 1. IOP Publishing, 2019.
18. Tao, Jing, et al. "Fetal birthweight prediction with measured data by a temporal machine learning method." BMC Medical Informatics and Decision Making 21 (2021): 1-10.

19. Emmanuel, Tlamelo, et al. "A survey on missing data in machine learning." Journal of Big data 8 (2021): 1-37.
20. Thaha, M. Mohammed, et al. "Advancing Environmental Sustainability through Implementation of Artificial Neural Networks for Wastewater Treatment Model Prediction and Remediation." Nanotechnology Perceptions (2024): 286-300.
21. Samad, Manar D., Sakib Abrar, and Norou Diawara. "Missing value estimation using clustering and deep learning within multiple imputation framework." Knowledge-based systems 249 (2022): 108968.