Handling Missing Values In Preterm Birth Dataset Using An Optimized Multinomial Naïve Bayes Model

Priyanka Garg[*1], **Dr. Sonali Goyal**[2] **and Dr. Ruby Bhatia**[3]

[1]Maharishi Markandeshwar (Deemed to be) University, Mullana, Ambala, Haryana (India).
[2]MMEC, Maharishi Markandeshwar (Deemed to be) University, Mullana, Ambala, Haryana (India)
[3]MMIMSR, Mullana, Ambala, Haryana (India)

Abstract—Missing values lower the quantity of information that machine learning (ML) methods learn during the training phase which harms classification accuracy. This research work presents an optimized multinomial Naïve Bayes (NB) classifier as a solution to this problem by handling the missing values. To lower the quantity of incorrect classification, it also suggests a feature selection and classification procedure. Our calculation of the research approach was implemented using the Kaggle dataset for maternal healthcare risk prediction. This research work compared the performance of different methods like decision tree (DT), and Random forest (RF). The outcomes of our simulations define that an optimized multinomial NB classifier attained the maximum accuracy rate of 94%. The DT classifier has achieved 66% accuracy and RF values attained 69%. The research method gives a reliable outcome for predicting maternal healthcare risk at PTB by identifying the problem of missing values in the dataset. The outcomes define that utilizing an optimized multinomial NB classifier improved the performance of the classification models. This work gives a reliable involvement in the domain of maternal healthcare risk prediction in Preterm Birth (PTB) research and the best part of identifying the missing values in ML uses

Index Terms—Preterm Birth, Maternal Healthcare Risk Prediction, Missing Values, ML, DT, RF.

I. INTRODUCTION

In the world, pregnancy-related issues and childbirth are the lives of 810 women per day. Maternal fatalities occur in 94% of countries with low or middle incomes [1]. While fewer women die giving birth because of technological advancements. Ensuring the safety of a pregnant woman and her unborn child remains a challenging task. Preventive actions and prognosis of difficulties might mitigate pregnancy-related risks in such a setting [2][3]. However, millions of moms' and babies' lives were saved when predictive modeling came into use. Protracted labour, Preeclampsia, and other obstetric z, labour that fails to advance (FTP), unusual fetal presentation, preterm birth, and other problems are among them [5]. Nonetheless, the bulk of these problems are not inevitable and the necessary actions are applied for the process of low-risk delivery.

A C-section or forceps birth, for instance, might be a safer delivery method, if the fetus presents abnormally [5]. Preterm birth (PTB) has been an important research analysis in the healthcare field for the last 20 years. Medical experts and scholars can now study efficient methods to mitigate the risk of PTB in women experiencing pregnancy-related difficulties, thanks to pregnancy and childbirth. Pregnant women are being provided with PTB control and education regarding early pregnancy warning symptoms through the implementation of healthcare services and medical measures [6]. Prenatal studies that provide specific clinical treatments for new-born's and new borns on their health, disease, care, etc.; outcomes heavily rely on the maternal history of the pregnant women. Newborns are pure souls and very special. They have no prior medical training, and their mothers' prenatal histories have a direct bearing on their early neonatal trajectory. Along with providing women with the necessary financial and emotional support before, during, and after pregnancy, the healthcare services also include learning, health, and other training programs that promote a healthy pregnancy. The common treatment of diseases is made by clinical experts depending on their experience.

In Nigeria, maternal mortality has been a major concern for women who live in rural and isolated positions. Long-distance travel requirements, a lack of primary healthcare specialists like paramedics and capable nurses, and other issues have created serious problems for the healthcare industry. Several cases of critical deprivation are experienced by healthcare managers, providers, and users, on the one hand, this prevents the managers from performing their duties, etc. It causes health issues and even mortality for the underprivileged population that needs healthcare services [7].

It's challenging to predict a pregnant woman's risk of PTB. There is only a 17–38% correlation between a diagnosis of PTB and the positive predictive values of true positives versus false positives [5].

In medical field, missing is a common problem and the hospitalized-based analysis is facing similar issues. The overall prognosis and individual's outcomes with missing data may fluctuate from those of those without missing data.

Existing research analyzing prenatal death predictors used easy methods like available or complete case analysis, hence ignoring significant data about missing information. In Statistical analysis, ignoring missing data frequently outcomes inaccurate and ineffective estimates of association, generally, when an information is missing at random [8].

Finding the optimal method for dealing with missing values in data that describe PTB was the aim of our proposed work. This dataset gathered at the Kaggle open-source site, was affected by the vast quality of missing attribute values. Furthermore, the Maternal Health Risk Dataset was inconsistent even though many attributes were numerical, which presents another challenge for data mining.

The main quality standard should be the sum of accuracy (ACC) conditional likelihood of full-term birth diagnosis and ACC (Prob.) of PTB diagnosis. The proposed work introduced an optimized Multinomial Naïve Bayes (NB) model (OMNBM) to handle the missing values. This proposed model is a combination of the Chi-square and NB methods [9]. The chi-squared feature selection method is used for feature selection and the multi-nominal NB classifier is used to handle the missing values. Among the several options to dealing with missing attribute values, the suggested method is among the easiest. This method used optimized multinomial NB trained on the considered values to classify the missing values. After that, it evaluates the performance metrics such as accuracy (ACC) rate and mean squared error (MSE).

The research article is organized as follows: Sect. 2 describes the prior work that has been carried out to handle the missing values. Sect. 3 discussed the research methodology of the proposed work. The simulation result analysis is defined in Sect. 4. Lastly, Sect. 6 discusses with conclusion and upcoming scopes.

II. PRIOR WORK

A. Preterm Birth (PTB)

Every year, more than 15 million newborns deliver early the 37th week of pregnancy. Around 12.9 million births globally in 2005 or about one-tenth of all babies born globally, premature births accounted for 9.6% of all births. Worldwide neonatal mortality and mortality rates are mostly caused by PTB, with notable variations in the global rate [10].

B. Types of Preterm Birth (PTB)

According to the gestational phase at birth, it has various classes. The phase counted from the first day of a female's last menstrual period (LMP) to the time of birth. This process is known as the gestational age. In general, it demonstrates that four kinds are covered:

(i) Under 28 weeks (Extreme PTB): It is the pregnancy.
(ii) 28-32 weeks (Very PTB): It is the delivery that occurs before the 28 and 32 weeks of pregnancy.
(iii) 32-34 weeks (Moderate PTB): It is the delivery that occurs before the 32-34 weeks of pregnancy.
(iv) 34-37 weeks (Late PTB): It is the delivery that occurs before the 34-37 weeks of pregnancy.

C. Existing Analysis

Several articles studied Machine learning (ML) models used to predict PTB risks, improving pregnancy outcomes by incorporating maternal characteristics from standard prenatal treatment, albeit at a time and expense. Although ML algorithms perform better than classical regression techniques, feature selection remains problematic because of insufficient reporting, confounding variables, improperly defined predictors, and unreliable data sources. Imeh Umoren et al. [7] drew attention to the problem of maternal death in rural Nigeria and proposed several solutions, including parametric-based maternal mortality, predictive decision-making, risk assessment systems, and machine learning-based decision tree models. Dr. Harish B. G. et al. [11] developed a predictive approach to lower maternal death rates and anticipate pregnancy issues using ML methods. The performance of the model reached 93% accuracy which was running on a Flask web server. According to Marc Hershey et al. [12] spontaneous preterm birth is a global public health concern. They emphasize the importance of precise prediction and identification of risk factors. ML model has achieved a false positive rate of 0.03. Xinxi Lu et al. [13] developed a bi-LSTM-based missing value imputation technique to address missing data in pregnancy examinations, aiming to prevent pregnancy-related hypertension and improve hospital management systems. Jerzy W et al., (2005) [14] determined the optimal approach for completing the missing attribute values in the Duke University preterm birth data. There were five different approaches employed: averages for numerical qualities, common values for symbolic attributes, and a unique MLEM2 method. The highest missing rates in numerical attribute values led to the conclusion that GMC-GA was the most efficient. Results from CMC-CA and CCF were poorer. Low-consistency data sets could not yield good results from MLEM2, which often generates fewer rules. The study concluded that while there is no one ideal technique for high-consistency datasets, some guidelines work better for low-consistency data sets. Alma B Pedersen et al. (2017) [15] described missing data as a common issue in clinical epidemiological research, affecting outcomes and prognosis. There are different types of missing shorts such as;

- Missing not at random (MNAR)
- Missing at random (MAR)
- Missing at random (MCAR)

In clinical epidemiological research, missing data was seldom MCAR, but it could pose challenges in analysis and interpretation, potentially weakening the results' validity. Under the MCAR assumption, statistical methods such as complete-case analysis, missing indicator techniques, single-value computation, and sensitivity analyses are used to derive unbiased estimates. Software for statistical analysis was subjected to the MAR assumption, which affected estimates for variables. It has not missing data as well as estimates for other variables. Janus Christian Jakobsen et al. (2017) [16] described missing data in randomized clinical trials could significantly impact inferences. The process of generating the data and the analytical techniques applied to correct it determined the possible bias resulting from missing the process generating the data and the analytical techniques applied to correct it determined the possible bias resulting from missing data. It was therefore necessary to arrange and pay closer attention to the analysis of trial data that had missing values. They discovered that managing missing data was a difficult undertaking and suggested analytical techniques to eliminate bias brought on by inevitable missing data [17]. They also talked about the benefits and drawbacks of multiple imputation, complete information maximum likelihood, and best-worst and worst-best sensitivity analyses. In addition, they offered useful flowcharts and a summary of the actions that must be taken into account, when a trial was analyzed.

D. Limitations in the Existing Medical Models

Nowadays, feature selection and classification methods are an important number of medical prediction methods that have been applied for better performance. Improving machine learning classifier performance and minimizing learning time are necessary because researchers find it difficult to select exact feature sets from maternal health datasets in linear time.

E. Main Contribution

Generally, optimized multinomial NB model based on the chi-squared metho d required for feature selection with multinomial NB method are introduced in this analysis to identify problems of the prior methods. The new method aims to enhance risk prediction accuracy and MSE rate by evaluating optimized features and classifying missing values from maternal health-related datasets.

III. Research Methodology

A. Objectives

The findings of the proposed work can be used to satisfy the leading objectives:

(i) Machine learning (ML) related missing values handle for PTB can be defined with the help of an optimized multinomial NB classifier model used to handle the missing values.

(ii) An introduced method is used to identify the feature selection and classify the missing attribute values.

B. Research Methodology

In this section, describes research methodology shown in Fig 1 as the working of the proposed framework. The proposed model steps, such as (i) Dataset uploading (ii) instance removal (iii) optimized Multinomial NB classifier (iv) Prediction, and (v) Evaluation.

It loads the dataset from a particular directory. It reads the dataset into a Panda data frame. It encodes the categorical target variable “risk_level” using a label encoder. It divides the dataset into training and testing sets. The test sets give instances with missing values in the “Month” column, while the train set gives instances with complete data. It SelectKBest applies to the feature selection process. Using with the chi2 (chi-squared) score function, the top 4 features are selected from the training data.

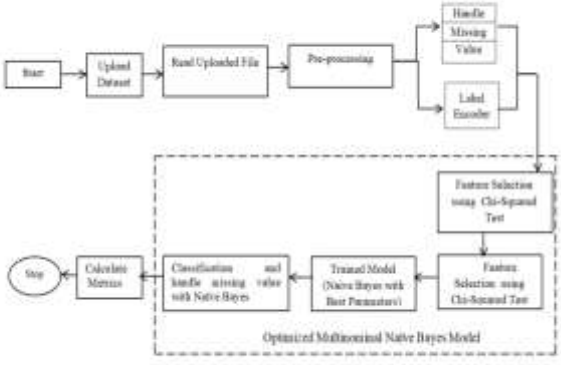
Start
Upload Dataset
Read Uploaded File
Pre-processing
Handle Missing Value
Label Encoder
Feature Selection using Chi-Squared Test
Feature Selection using Chi-Squared Test
Trained Model (Naive Bayes with Best Parameters)
Classification and handle missing value with Native Bayes
Calculate Metrics
Stop
Optimized Multinomial Native Bayes Model

Fig. 1. Flowchart of Proposed Methodology

This technique enhances model performance and training time by reducing the dimensionality of the data. Here, an optimized Multinomial (Naive Bayes) model is used. RandomizedSearchCV is employed to find the best hyperparameters for the model. The param_distributions dictionary specifies the range for the alpha parameter (smoothing parameter for MultinomialNB) to be searched. n_iter=50 means 50 different combinations of hyperparameters will be tried, and cv=5 specifies 5-fold cross-validation. The best hyperparameters and the corresponding best score are printed. The model with the best parameters (best_model) is used to make predictions on the test data. Model Training and Prediction: It initializes a Multinomial Naive Bayes model. Trains the model using the training data. Predicts the missing values for the Month; column in the testing set using the trained model. Calculates the accuracy of the predictions using the Jaccard similarity score. Accuracy Comparison: Loads pre-calculated accuracies of other models (Decision Tree and Random Forest) from pickled files. Compares the accuracy of the Optimized Multinomial NB model with the accuracies of the other models. The detailed steps are defined as follows:

DATA COLLECTION

This section describes the dataset in detail. Information was gathered from a number of hospitals, public health centre, and protective health care facilities using an Internet of Things-based risk monitoring system (RMS) [18]. Here are some attributes are follows as:

- **SystolicBP** means an Upper level of BP in mmHg unit, another important attribute value for the period of pregnancy.
- **Diastolic BP** means a lower level of BP in mmHg unit, another important attribute value for the period of pregnancy.
- **Blood Sugar** means glucose levels in the form of molar concentration, mmol per L.
- **Risk Level** means predicted risk intensity level for the duration of pregnancy considering the prior attribute.
- **Weight (Wt.)** means wt. gain from the prior month in Kilograms.
- **Month** means the current month of the pregnancy.
- **Trimester** means the current trimester of the pregnancy.

Above mentioned attributes are utilized as input variables and the result is the risk level in PTB defined by the status as High (H), Low (Low), and No Risk (NR).

DATA PRE-PROCESSING

It is the procedure of removing an accurate record from an original dataset. It identifies the null values and uses the label encoder step. The label encoder is a process that is used to convert alphabetical data values into integer format. After this process, it generates the unique values format. After that, it uses the isnull ().sum() to calculate the null values. Table I shows the preprocessing outcome.

TABLE I PREPROCESSING OUTCOME

Sn o.	SystolicBP	DiastolicBP	Blood Sugar	Wt.	Risk Level
1	110	80	56	72	0
2	95	60	76	78	1
3	110	80	70	200	2
4	100	80	87	75	3
5	110	80	86	85	1

Feature Selection Using Chi-Square Test

In ML, the chi-square test [19] is frequently utilized for feature selection (FS). The main motive of FS is to choose a subset of feature sets that are reliable to handle the missing values. It is utilized in statistics to analyze the independence of events. Suppose the data values of two variables, such as (1) observed count 0 and (ii) expected count E. The formula is used as follows:

$X_c^2 = \Sigma (O_i - E_i)^2 / E_i$ .... (1)

Here eq (1), which has degrees of freedom (c), observed values (O), and expected values (E), illustrates the relationship between the variables.

Generally, three-factor clusters are chosen as the most significant features for handling the missing values in PTB. Attributes are upper-level BP (systolic BP), Weight, and trimester. All the missing values are handled and variables that were derived from the questionaries' are mentioned in Table II.

TABLE II SELECTED CHI-SQUARE TEST

Attributes	Levels
SystolicBP	2 (predictor)
Wt	4 (predictor)
Trimester	5 (predictor)

MODEL FORMULATION

The Maternal health dataset is divided based on the criteria of decision tree (DT), RF, and Optimized Multinomial Naïve Bayes (NB). The methods are used on each class of dataset to calculate their performance.

Generally, a simple NB method is used for classification. Optimized Multinomial NB kind of classification used for the research work. Typically, the dataset under the classification method is separated into two sections. 1/3rd of the dataset is utilized as test data and the left data is utilized as train data. The connection between the values of the predictors and the target class is defined in the training procedure of the classification method. The predicted values are related to the target_values in a set of information in the classification methods. NB is a category of supervised learning (SL) method based on Bayes' theorem. It accomplishes so by utilizing the independent predictor hypothesis. The conditional probability for the subsequent classification of the research effort is the basis of the Bayes theorem. The statistical method for classifying the dataset is defined by the Bayesian classification.

To predict the text labels two theories are applied such as probability theory and the Bayes theorem. The NB classifier is used for accurate data categorization.

OPTIMIZED MULTINOMIAL NAÏVE BAYES CLASSIFIER

This classifier is the initial SL method that was introduced. It is a probabilistic approach to learning. It supplies the multinomial divided data required for the NB method. This method is reliable for classification with discrete features. This method is utilized when the data is separated multinomial that is multiple occurrences matter more. This division is parameterized by vectors $\theta_Y = (\theta_{Y1,} ...., \theta_{Yn})$ for individual class Y, where n represents the no. of features, and θ_{Yi} is the prob. P'(X_i | Y') of feature i looking at a method relating to the Y class.

The subsequent research implemented method is an optimized multinomial NB method. This method works to foretell the Maternal Health risk prediction between pregnant women. The initial step in the optimized method is to apply the Chi-squared test to the dataset containing several maternal health indicators such as age, BP, etc. The test verifies the most important indicators that are strongly connected with pregnancy-based risks like gestational diabetes, etc. Once the most reliable feature sets are selected the multinomial NB classifier is applied. The main advantages of the optimized Multinomial NB classifier are follows as: (i) Enhanced accuracy rate: it focuses on the most important feature sets, and the model can more precisely predict maternal health risks. (ii) Efficiency means reducing the number of features to lead to faster training and prediction times. (iii) The chosen feature sets given correspond to well-defined risk factors, making the model's predictions more interpretable for healthcare professionals. An optimized multinomial NB classifier is the simplest method that can efficiently address and use the most reliable risk factors, leading to more precise and actionable predictions. It is especially reliable in pregnancy, where timely and precise risk assessments can more significant difference in results for both the mother and baby.

DECISION TREE (DT)

It [19] is a flexible supervised learning method that is used in regression and classification applications. Its name comes from the way it looks, which is like a tree structure with branches. Recursive data division is used by the decision tree classifier, a predictive model, to generate predictions or judgments. It traverses the tree and categorizes new instances according to patterns it has learned using a sequence of if-else conditions. Because the decision-making process is so clearly understood and depicted, this model is renowned for its interpretability. Fig 2 shows an example of DT.

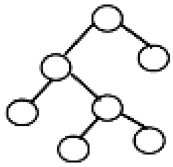

Fig. 2. Decision Tree Architecture [11]

RANDOM FOREST (RF)

It is a regression tool that generates predictions by combining many decision trees using an ensemble approach. A randomized subset of training data and input attributes are used by the RF classification algorithm Regressor, a decision tree ensemble, to train each tree shown in Fig 3.

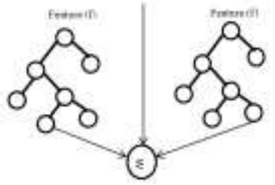

Fig. 3. Illustration of RF [11]

The accuracy (ACC) rate and mean squared error (MSE) rate were used to calculate the performance of the model.

IV. EXPERIMENTS

It is recommended that footnotes be avoided (except for the unnumbered footnote with the receipt date on the first page). Instead, try to integrate the footnote information into the text and the reference part.
The experiment was analyzed using PYTHON and Scikit-Learn Library for the maternal health risk prediction dataset with 1014 patient observations, and PANDAS library for "*.csv" file processing.

A. Performance Evaluation
This section discussed the performance evaluation such as ACC and MSE rates.

Accuracy (ACC)
The proportionality of exact categorized values is known as accuracy. The total of true positives (T_p) and true negatives (T_n) is compared to the total of T_p, false positives (F_p), T_n, and false negatives (F_n) in order to statistically determine it. It is defined in eq (2).

$$Accuracy = \frac{T_p+T\_n}{T_p+T_n+F_n+F\_p} \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Mean Square Error (MSE)

The average of absolute error values of positive numbers is used to determine the MSE score. An MSE score that is positive is guaranteed when there is a positive (yi) difference between the expected and forecasted values $\widehat{y_i}$. It is described in eq (3).

MSE = $\frac{1}{n}\sum_{i=1}^{n}|y_i - y'_i|$.............(3)

B. Discussion of Results

Testing the Maternal Healthcare Risk Prediction Dataset with both missing values outcomes in overfitting; however, when missing values were imputed, overfitting did not occur. The optimized Multinomial NB model can manage the missing values. In the start, when comparing ACC and MSE metrics for the training set with missing values, the performance has been improved as compared with the existing methods. The implemented technique to resolve the missing values is to eliminate the incomplete. Table III shows the proposed model used metrics. The RandomSearchCV range is defined to improve the accuracy rate. Iterations are used in the proposed model value of 50. Cross-validation means calculating the estimator performance and value set of 5 and the random state value is 42.

TABLE III THE PROPOSED MODEL USES METRICS

Metrics	Values
Randomized SearchCV range	0.1-2.0
Iterations	50
Cross-validation	5
Scoring	Accuracy and MSE
Random State	42

The following outcome shows the ACC accurately addresses maternal healthcare risk for the defined dataset of metrics for several classification methods. These ML methods are considered, and the optimized multinomial NB yields the maximum ACC for precise classification in both training set data. The proposed optimized multinomial NB classifier achieves the maximum ACC of 94% when the missing values are managed. The decision tree attains a minimum value of 66%. Several ML methods for training and testing, it is considered optimized multinomial NB classifier that displays importantly maximum efficiency and ACC compared to other methods. When the number of missing values is more then it means a large negative performance in classification. This problem is managed by the optimized NB classifier. The proposed model's outcomes shown in Table IV with several classification models for maternal risk prediction. Fig 4 shows the ACC and MSE comparison of all the ML methods.

TABLE IV PROPOSED MODEL OUTCOMES

METHODS	Accuracy (%)	MSE (%)
Optimized Multinomial NB model	94.99	5.01
Decision Tree	66.66	33.4
Random Forest	69.56	30.44

The proposed optimized multinomial NB classifier is managing the missing values that have improved the performance of classification. The maximum ACC was achieved by the optimized multinomial NB classifier with 94%. RF attains 69% ACC whereas the DT attains the minimum ACC of 66%. Table IV defines the ACC and MSE of the ML methods when the research methodology is utilized.

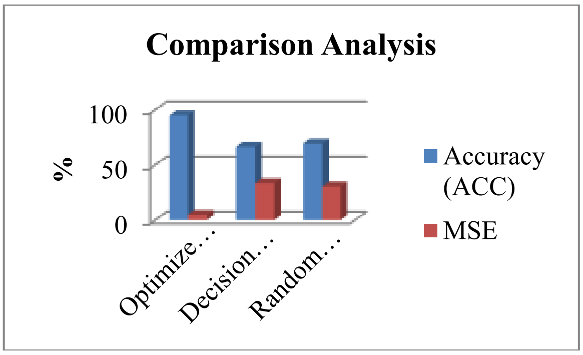

Fig. 4. Accuracy comparison of different ML methods.

C. Limitations

The proposed limitations of this analysis are as follows:

- The research classifier may be computationally difficult, normally when handling huge amount of datasets with several missing values.
- The implemented method depends on the training of an optimized multinomial NB classifier for each missing value Overfitting may occur if the methods are too difficult or if the dataset is small.

V. Conclusion

This research article defined a technique for managing missing values in an optimized multinomial NB classifier, DT, and RF for maternal healthcare risk prediction. The objective of this research was to lower the difficulties involved in applying dependable data handling techniques while simultaneously raising the total outcome accuracy rate. The research technique utilizes an optimized multinomial NB classifier to handle the missing values. The outcomes of the analysis defined that the researched approach attained enhanced accuracy and mitigated the error rate when compared to the other ML methods. The highest accuracy of 94% was achieved using an optimized multinomial NB classifier, which demonstrated the efficiency of the implemented model. The implemented model has numerous benefits over predictable techniques for handling missing values.

The conclusion of this research work is the implementation of an optimized multinomial NB classifier for handling missing values in ML methods for maternal healthcare risk prediction has evaluated capable outcomes. This work has the potential to be used for various categorization tasks and offers insightful data about the subject of ML.

Appendix

Sr No.	Abbreviations	Descriptions
1.	ML	Machine Learning
2.	DT	Decision Tree
3.	RF	Random Forest
4.	PTB	Preterm Birth
5.	NB	Naïve Bayes
6.	ACC	Accuracy
7.	MSE	Mean Square Error
8.	LMP	Last Menstrual Period
9.	MNAR	Missing not at random`
10.	MAR	Missing at random
11.	MCAR	Missing at random
12.	RMS	Risk Monitoring System
13.	H	High
14.	L	Low
15.	NR	No Risk

16.	SL	Supervised Learning

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data Availability Statement

In this research work, we were using an open-access dataset that was gathered by the Kaggle site. This dataset is used for research work and is freely accessible in the Kaggle.

REFERENCES

[1] Trends in maternal mortality 2000 to 2017: estimates by who, unicef, unfpa, world bank group and the united nations population division2000-2017. Available:https://www.unfpa.org/featured-publication/trends-maternal-mortality
[2] Maternal deaths decline slowly with vast inequalities worldwide. World Health Organ. Available: https://www.who.int/news/item/19-09-2019-maternaldeaths-decline-slowly-with-vast-inequalities-worldwide. Maternal deaths decline slowly with vast inequalities worldwide. World Health Organ. https://www.who.int/news/item/19-09-2019-maternaldeaths-decline-slowly-with-vast-inequalities-worldwide.
[3] Maternal mortality. Available:https://www.who.int/news-room/fact-sheets/detail/ maternal-mortality.
[4] P. Agrawal "Maternal mortality and morbidity in the United States of America." Bulletin of the World Health Organization. 2015;93:135.
[5] M. N. Islam, S. N. Mustafina, T. Mahmud and N. I. Khan. "Machine learning to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda." BMC pregnancy and childbirth. 2022, vol 22.
[6] R. Raja I. Mukherjee, and B. K. Sarkar, "A Machine Learning-Based Prediction Model for Preterm Birth in Rural India. Journal of Healthcare Engineering. 2021;2021(1):6665573.
[7] I. Umoren, A. Silas, and B. Ekong B. "Modeling and prediction of pregnancy risk for efficient birth outcomes using decision tree classification and regression model." InArtificial intelligence and soft computing 21st international conference, ICAISC 2022 pp. 19-23.
[8] I. B. Mboya, M. J. Mahande, J. Obure, and H. G. Mwambi. "Predictors of perinatal death in the presence of missing data: A birth registry-based study in northern Tanzania." Plos one. 2020 vol 15, no 4:e0231636.
[9] S. Beck, D. Wojdyla, L. Say, A. P. Betran, J. H. Requejo, C. Rubens R. Menon, and P. F. Van Look. "The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity." Bulletin of the world health organization. 2010;88:31-8.
[10] R. Pari, M. Sandhya, and S. Sankar. "Risk factors based classification for accurate prediction of the Preterm Birth." In2017 International Conference on Inventive Computing and Informatics (ICICI) 2017 Nov 23, IEEE, pp. 394-399.
[11] B. G. Harish, K. G. S. Chetan, N. Sushma, B. N. R. Sabeeha, N. B. Priyanka, and M. N. Shreyanka. "Pregnancy Risk and Fetal Health Prediction using Machine Learning" 2023, vol 10, no. 7, pp-14-121.
[12] M. Hershey, H. H. Burris, D. Cereceda, and C. Nataraj. "Predicting the risk of spontaneous premature births using clinical data and machine learning." Informatics in Medicine Unlocked. 2022 ,1;32:101053.
[13] X. Lu, L. Yuan, R. Li, Z. Xing, N. Yao, and Y. Yu. "An improved Bi-LSTM-based missing value imputation approach for pregnancy examination data." Algorithms. 2022 vol 16, no. 1.
[14] J. W. Grzymala-Busse, L. K. Goodwin, W. J. Grzymala-Busse, and X. Zheng. "Handling missing attribute values in preterm birth data sets." InRough Sets, Fuzzy Sets, Data Mining, and Granular Computing: 10th International Conference, RSFDGrC 2005, Regina, Canada, August 31-September 3, 2005, Proceedings, Part II 10 2005, pp. 342-351. Springer Berlin Heidelberg.
[15] A. B. Pedersen, E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, and I. Petersen. "Missing data and multiple imputation in clinical epidemiological research." Clinical epidemiology. 2017, pp-157-66.

[16] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. BMC medical research methodology. 2017 Dec;17:1-0.

[17] M. Ahmed, M. A. Kashem, M. Rahman, and S. Khatun. "Review and analysis of risk factor of maternal health in remote area using the Internet of Things (IoT)." InInECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 2019-2020, pp. 357-365. Springer Singapore.

[18] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao. "A chi-square statistics based feature selection method in text classification." In2018 IEEE 9th International conference on software engineering and service science (ICSESS) 2018, pp. 160-163. IEEE.

[19] P. Patidar, and A. Tiwari, "Handling missing value in decision tree algorithm." International Journal of Computer Applications. 2013, vol 13.

Ms. Priyanka Garg is a research scholar at Department of computer Science & Engineering in MMEC, Maharishi Markandeshwar (Deemed to be) University, Mullana.Pursuing PhD in the field of Machine Learning. During 11 years of her academic career she has been actively engaged in teaching, mentoring and conducting research. Have authored 7 research papers in various aspects of computer science.
Scholarpriyanka85@gmail.com

Dr. Sonali Goyal is an accomplished academic and researcher in the field of machine learning and the Internet of Things. As an associate professor in Department of CSE, MMEC, Maharishi Markandeshwar (Deemed to be) University, Mullana. In terms of research, Dr. Sonali Goyal has a prolific record, having authored 25 research papers on various aspects of machine learning and the Internet of Things. Her work demonstrates a deep understanding of the underlying principles and a keen ability to apply them to practical problems. In addition to her research papers, she has also obtained five patents for her innovative contributions to the field. She has also authored a book, showcasing her ability to communicate complex concepts effectively.
sonaligoyal@mmumullana,org

Dr. Ruby Bhatia, M.D FICOG , Professor and Head ,MMIMSR, Mullana, Ambala is a Master trainer for HPV Vaccination Project Haryana- FOGSI and President NIGF for Haryana state. Chaired sessions & panelist in various national and state FOGSI conference. Examiner and Paper setter MBBS/M.D. at AIIMS New Delhi and PGIMS Rohtak, MMIMSR, Baba Farid University, PB. 85 research publications in National and International indexed journals
Hod.obsgynae@mmumullana.org