Maternal Health Analysis

Mayuri Kothari, Ritunjaya Singh Thakur, Anjasha Singh
210911368, 210911358, 210911244

April 2024

Abstract

Maternal mortality due to complications during and after pregnancy remains a significant public health concern, particularly in rural areas where access to adequate healthcare services is limited. Insufficient knowledge about maternal healthcare exacerbates this issue, leading to preventable deaths of pregnant women and newborns. Early monitoring and intervention during pregnancy are crucial for ensuring the healthy development of the fetus and safe delivery. In this study, we investigate various factors that impact the health of pregnant women and aim to develop predictive models to assess the level of health risk during pregnancy. We employ three machine-learning algorithms: Random Forest, Naive Bayes, and Logistic Regression, to analyze a dataset comprising demographic, medical, and socioeconomic variables. By leveraging these models, we aim to provide healthcare practitioners with a tool for early identification of pregnant women at higher risk, enabling targeted interventions and ultimately reducing maternal mortality rates. The findings from this study can inform policymakers and healthcare professionals in designing effective strategies to improve maternal healthcare and mitigate the risk of maternal mortality and morbidity.

1 Introduction

Every pregnancy is a profound journey, filled with hopes, dreams, and anticipation. Yet, for too many women worldwide, this journey is fraught with peril, overshadowed by the silent epidemic of maternal mortality. Despite advancements in medical science, hundreds of thousands of women succumb to pregnancy-related complications each year, underscoring the stark inequalities and inadequacies entrenched within global healthcare systems.

Nowhere is this disparity more evident than in underserved regions, where access to quality healthcare remains a distant aspiration. Here, the promise of safe motherhood often remains elusive, as women navigate the treacherous terrain of pregnancy without adequate support or resources. It is within these communities that the devastating impact of maternal mortality is most keenly felt, perpetuating a cycle of despair and inequality.

At the heart of this crisis lies a fundamental truth: the lack of sufficient knowledge about maternal healthcare during and after pregnancy serves as a formidable barrier in the fight against maternal mortality. Inadequate monitoring, coupled with untimely interventions, compounds the risks faced by both mothers and newborns, leaving them vulnerable to life-threatening complications that could otherwise be prevented.

Indeed, the journey of pregnancy, from conception to childbirth, demands unwavering vigilance and care. It is a delicate dance between the health of the mother and the well-being of the unborn child, a dance made all the more challenging in the absence of accessible healthcare services. In rural areas especially, where infrastructure and resources are scarce, this dance becomes a harrowing ordeal, with women forced to traverse the perils of pregnancy alone.

Acknowledging these profound challenges, this notebook embarks on a mission to confront the gaps in maternal healthcare with determination and resolve. Through the transformative power of machine learning, we endeavor to predict the level of health risk for pregnant women, arming healthcare providers with timely and actionable insights to guide their interventions.

Through a comprehensive examination of key factors influencing maternal health—including age, blood pressure, blood glucose levels, body temperature, and heart rate—we seek to develop a robust predictive model capable of assessing risk intensity levels throughout pregnancy. By harnessing the vast potential of data analytics, we aim to shed light on the intricate interplay between these variables, unraveling the mysteries of maternal health and mortality.

Our ultimate objective is clear: to equip healthcare practitioners with a potent tool for early identification of pregnant women at higher risk, enabling them to deliver targeted interventions and personalized care plans. In doing so, we aspire to rewrite the narrative of maternal healthcare, transforming it from one of despair to one of hope and empowerment.

The promise of our solution is profound. By harnessing the synergistic capabilities of machine learning algorithms and comprehensive data analysis, we endeavor to bridge the chasm in maternal healthcare knowledge, paving the way for improved outcomes for both mothers and newborns. Through the transformative potential of predictive modeling, we aim not only to save lives but also to champion the cause of maternal and child health on a global scale, ushering in a new era of proactive and personalized care delivery.

1.1 Objectives:

- The primary goal is to construct a machine learning model capable of forecasting the health risk level for pregnant women based on a range of factors such as age, blood pressure, blood glucose levels, body temperature, and heart rate. This model will utilize historical data to discern patterns and correlations between predictor variables and the target variable (risk level), facilitating accurate assessment of health risks associated with pregnancy.

- Identifying pivotal factors influencing maternal health constitutes another key objective. Through thorough data analysis, the notebook aims to pinpoint and examine factors significantly impacting maternal health during pregnancy. By unraveling the connections between these factors and the risk level, healthcare practitioners can glean valuable insights into the determinants of maternal health outcomes, thereby tailoring interventions accordingly.

- Enabling early detection of high-risk pregnancies stands as a crucial aim. The notebook endeavors to furnish healthcare providers with a tool for promptly identifying pregnant women at elevated risk of complications during pregnancy. By leveraging the predictive model developed herein, healthcare practitioners can proactively intervene and devise personalized care plans to mitigate the risks associated with maternal health complications, ultimately fostering improvements in maternal and neonatal outcomes.

- Enhancing maternal healthcare in underserved communities represents another vital facet. By bridging gaps in maternal healthcare via predictive modeling, the notebook seeks to contribute to the amelioration of maternal healthcare services, particularly in underserved locales where access to quality healthcare is scant. By furnishing healthcare practitioners with actionable insights and tools for risk assessment, the objective is to diminish disparities in maternal health outcomes and advocate for equitable access to healthcare services for pregnant women.

- Informing policy and decision-making serves as the final overarching objective. The project endeavors to generate insights capable of guiding policy and decision-making in the realm of maternal healthcare. By discerning trends, patterns, and risk factors linked to maternal mortality and morbidity, policymakers can formulate targeted strategies and allocate resources effectively to address the underlying causes of maternal health disparities, thereby fostering enhancements in overall maternal and neonatal health outcomes.

2 Literature Review

The study detailed in [1] conducted a thorough examination of the utilization of artificial intelligence (AI) and affective computing (AC) in the domain of pregnancy health and well-being. It identified a burgeoning interest in employing AI for monitoring health during pregnancy, enhancing decision-making processes, and mitigating maternal-foetal morbidity. Various algorithms, including decision trees, support vector machines, logistic regression, and artificial neural networks, were commonly applied. Model performance was typically evaluated using metrics such as accuracy, recall, and area under the receiver operating characteristic curve. The review underscored research gaps, notably the limited exploration of AC to support psychological health during pregnancy, and raised concerns regarding data security and privacy. Overall, the review offers valuable insights for researchers and practitioners keen on developing AI-based solutions for obstetrics and gynaecology.

The research discussed in [2] aimed to investigate the determinants of Caesarean sections (C-sections) and employ machine learning methods for classifying birth data in Muzaffarabad, Pakistan. By collecting data from two government hospitals and employing dimensionality reduction techniques, the study analyzed 23 factors with a sample size of 488 subjects. Various classification methods were employed, including Random Forest, Linear Discriminant Analysis, Support Vector Machine, Naive Bayes, K-nearest neighbors, Adaboost, and Neural networks. The study concluded that Random Forest exhibited the highest accuracy, reaching 91.8%, suggesting its potential for developing decision support systems to predict birth modes, thereby contributing to women's health and childcare.

In [3], researchers explored the application of machine learning techniques for the early diagnosis of hypertensive disorders in pregnancy, a crucial aspect of women's healthcare. They proposed using algorithms such as Naïve Bayes, decision tree induction, support vector machines, and artificial neural networks to analyze data from pregnant women with hypertensive disorders. Evaluation metrics included true positive rate, false positive rate, and receiver operating characteristic curves. The study identified the averaged one-dependence estimators (AODE) algorithm as the top performer, showcasing its potential for enhancing healthcare decision-making in high-risk pregnancies. The conclusion emphasized the significance of smart decision support systems in personalized healthcare and advocated for further research to optimize and apply these techniques to other gestational diseases.

The researchers in [4] utilized computational intelligence techniques, including artificial neural networks (ANNs), support vector machines (SVM), and k-nearest neighbors (k-NN), for predicting aneuploidy risk. Data collected from pregnant women attending foetal medicine centers were analyzed, incorporating parameters such as maternal age, previous pregnancy history, ultrasound measurements, and biochemical results. Statistical analysis was performed to assess separability between feature distributions and test for normality. The study found that the ANN-based diagnostic system exhibited promising results, particularly in detecting T21 pregnancies with low false positive rates. While SVM and k-NN also yielded results, ANNs outperformed them. The study concluded that the proposed system could effectively predict aneuploidies, including T21, with potential clinical applicability. Future research directions include exploring alternative neural network architectures and integrating additional parameters to enhance prediction accuracy, such as paternal information.

In [5], researchers proposed an integrated solution for continuous monitoring of high-risk pregnancies using IoT sensors, fog computing, big data analytics, and a convolutional neural network (CNN) classifier. The system included an emergency diagnostic subsystem based on fixed thresholds and inference rules, achieving accuracies exceeding 90% for foetal, maternal, and combined emergencies. Feature extraction modules extracted 15 parameters from foetal heart rate processing and 30 parameters from maternal vital signs monitoring. CNN prediction subsystems were evaluated, with a 1-D CNN with six convolutional layers and 10-second windows demonstrating the best performance for classifying maternal and foetal health status. Overall, the integrated solution showed promising results for ambulatory maternal and foetal monitoring, with potential for clinical application and further research.

In [6], Anyi Cheng focused on improving labour and preterm delivery prediction through Electro hysterogram (EHG) analysis. The research aimed to enhance prediction accuracy by employing a machine learning framework that combines feature extraction, selection, and classification algorithms. Notably, the study introduced novel features such as cross and multichannel entropy alongside traditional features like Mutual Information (MI) and spectral features. Utilizing Support Vector Machines (SVM) as the primary classification algorithm, the study evaluated performance using metrics including accuracy, sensitivity (SN), specificity (SP), and Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) curves. Results demonstrated notable advancements in prediction accuracy, with accuracies ranging from 90.5% to 96.4% for labour and preterm prediction, respectively, surpassing prior literature benchmarks. The study's conclusions underscored the importance of feature selection strategies, the impact of utilizing contraction segments versus the entire EHG signal, and the potential clinical relevance of automated contraction segment detection algorithms. Additionally, the study highlighted the significance of incorporating novel features like cross and multichannel entropy to better capture the spatiotemporal regularity of uterine contractions, thereby improving prediction outcomes, and enhancing obstetric care.

While in [7], although the objective remains consistent, the methodology involved collecting and analyzing existing research on high-risk pregnancy complications, including Aortic segmentation, Preeclampsia, foetal growth restriction, Preterm birth, Anaemia, and Gestational diabetes mellitus. The challenges in predicting high-risk pregnancies were identified, such as limited development in resource-poor countries, lack of awareness and access to healthcare, and difficulty in predicting specific risks. The authors reviewed multiple machine learning methods for risk prediction, including Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes, and Random Forest. Results showed that the Decision Tree method achieved the highest accuracy rate of 93.70% compared to SVM (90%), KNN (86%), and Naive Bayes (75.18%). Consequently, it was concluded that machine learning models can effectively predict pregnancy risks, with Decision Tree exhibiting the best performance among the analyzed methods, thus highlighting the potential of such models to improve maternal health outcomes.

The objective of the research in [8] is to mitigate the mortality rate for pregnant women by implementing preventive measures as early as possible for those identified as high-risk during pregnancy. The methodology involved a data analysis approach, where data were collected from 20 samples of pregnant women meeting specific criteria. The analysis encompassed three steps: description analysis, predictive analysis, and prescription analysis. The study concluded by categorizing the risk level of survival into three categories based on identified risk factors, thereby offering tailored recommendations for pregnant women at different risk levels.

In [9], a recent analysis delved into expert systems for pregnancy risk detection, uncovering key trends through a systematic review of existing literature. Rule-based systems emerged as a common approach, valued for their simplicity and ease of implementation, albeit reliant on expert knowledge. Fuzzy expert systems were identified as another viable option, capable of handling ambiguity and uncertainty in healthcare contexts. Artificial neural networks (ANNs) showcased high accuracy potential but demanded large datasets for training and lacked interpretability. While other machine learning techniques offered diverse methods, their reliability varied and required further refinement for consistent performance. The study identified hypertension, premature birth, pregnancy abnormalities, and ectopic pregnancy as primary risks detected by these systems, with performance metrics focusing on accuracy, sensitivity, and specificity. Conclusively, expert systems held substantial promise for pregnancy risk detection, with rule-based and ANNs currently prevailing, each presenting distinct advantages and challenges.

This study [10] proposed a blockchain-based maternal health information system integrated with predictive analytics to securely store and share health data, ensuring privacy and integrity. Participants were categorized as patients, doctors, community health workers, and moderators, with smart contracts managing data access and interactions. Healthcare workers recorded patient data on the blockchain, while doctors verified and approved entries. The system employed a Random Forest classifier to predict pregnancy complications, achieving an accuracy of 83.32%. The authors concluded that integrating blockchain and machine learning offers a promising solution for secure data management and risk prediction in maternal health.

In [11], the study tackled maternal mortality in low-income regions like Bangladesh by developing an IoT-based maternal healthcare system. Wearable sensing devices and questionnaire-based data collection were integrated to gather comprehensive health information. Machine learning algorithms, notably decision trees, were utilized for risk classification, achieving a remarkable 97% accuracy. Blood sugar emerged as a critical risk factor during pregnancy. The study proposed a website for backend risk analysis and suggested future research to expand attribute sets and leverage big data for enhanced predictive accuracy, thereby offering a promising pathway to address maternal mortality rates.

The research in [12] aimed to predict maternal health risk factors using machine learning models. A dataset of maternal health risk factors was collected and preprocessed, and five machine learning models were employed for prediction. The Cat Boost model exhibited the best performance, with an accuracy of 89.5%, precision of 90.2%, recall of 88.9%, and F1-score of 89.5%. The authors concluded that machine learning models could effectively predict the risk level of pregnant women, thereby aiding in proactive healthcare management.

In [13], Lyu and Liang addressed the task of predicting pregnancy outcomes during the COVID-19 pandemic using electronic health records (EHR) and machine learning techniques. Their algorithm, Temporal Events Detector for Pregnancy Care (TED-PC), utilized a rule-based approach to identify clinical events throughout gestational weeks. Supervised machine learning models were employed, tailored to capture temporal dynamics in pregnancy. Preliminary results showed promise in exploring the association between clinical outcomes and risk factors, paving the way for enhanced clinical decision support systems and improved maternal health management.

The objective of the research in [14] was to predict the likelihood of pregnancy within a menstrual cycle based on data collected during the first 24 days of that cycle. Four predictive models were explored, including logistic regression, Long Short-Term Memory (LSTM) networks, LSTM combined with the Barrett-Marshall and Schwartz (BMS) fertility model, and LSTM with user embeddings. Model performance was evaluated using the Area Under the Curve (AUC) metric, with LSTM with user embeddings achieving the highest AUC of 0.67. The study also investigated model interpretability, revealing consistent associations between sexual activity during the fertile window and pregnancy outcomes, consistent with prior fertility research.

Finally, the research in [15] aimed to identify empirical practices in treating pregnant patients with and without urinary tract infections (UTIs) using data analysis of medical records. The UTIW (Urinary Tract Infection Workflow) was employed, involving domain-specific spell checking, text normalization, and rule extraction. The study provided insights into UTI diagnosis and treatment practices during pregnancy, highlighting the potential of data analysis techniques to improve maternal-foetal health outcomes by identifying patterns in medical records and informing clinical decision-making.

3 Methodology

1. Data Collection and Preprocessing: The dataset, containing seven variables including age, systolic blood pressure (SystolicBP), diastolic blood pressure (DiastolicBP), blood glucose levels (BS), body temperature (BodyTemp), heart rate (HeartRate), and risk level, was obtained from credible sources. Preprocessing steps involved cleansing the dataset to address missing values, outliers, and inconsistencies. Additionally, categorical variables underwent encoding, while feature scaling was applied to ensure uniformity across all variables.

2. Feature Selection: Pre-modeling efforts included feature selection techniques such as correlation analysis and assessment of feature importance. This process aimed to identify the most pertinent predictors for the target variable (risk level), optimizing model performance by focusing on the most informative features while mitigating computational complexity.

3. Model Development: Three machine learning algorithms—Random Forest, Logistic Regression, and Naive Bayes—were chosen for model development due to their suitability for classification tasks and ability to handle both numerical and categorical data. Each model underwent training on the preprocessed dataset, utilizing a portion of the data for training and the remainder for validation to assess performance.

4. Model evaluation: Model performance indicators such as accuracy, precision, recall, F1 score, and area under the ROC curve are used to evaluate. Cross-validation methods such as k-fold cross-validation ensure the robustness and generality of the model. Comparing each performance model can identify its strengths and limitations and suggest model options and improvements.

5. Interpretation and Communication of Results: The final phase involved interpreting model outcomes and communicating findings to stakeholders such as healthcare professionals, policymakers, and community members. Insights derived from the models were leveraged to inform clinical practice, policy formulation, and future research agendas, with the overarching aim of enhancing maternal healthcare outcomes and diminishing maternal mortality rates.

3.1 Random Forest

Random forest is a widely used machine learning algorithm belonging to the ensemble learning family and is an important tool for maternal health assessment and classification. Its versatility allows it to adeptly perform classification and regression tasks, making it an ideal candidate for predicting important outcomes in maternal health. In this context, random forest evidence is invaluable in assessing the risks associated with pregnancy problems that can cause preterm birth and predisposition to many conditions. Random Forest uses a variety of factors, including demographic data, medical history, and lifestyle, to create predictive models that allow doctors to identify and time high-risk pregnancies. In addition, factor analysis supports the interpretation of data and provides a better understanding of decisions regarding maternal health. This valuable information not only informs clinical decision-making but also helps develop effective public health policies to promote nutrition services. In essence, the forest is a powerful ally in the quest to reduce maternal health risk by providing strong, accurate and interpretive support in parental care.

Steps of the methodology for implementing a Random Forest model for maternal health risk analysis and classification:

1. Data Preprocessing:
    - Outlier Handling: Outliers, which are data points significantly different from other observations, can skew the model's learning process. In this step, outliers, such as records with unrealistic values (e.g., a heart rate of 7 bpm), are identified and handled appropriately. Handling methods may include dropping these records or applying transformations to mitigate their impact.
    - Feature selection: Not all features in the dataset contribute equally to predicting the target variable. Irrelevant or repetitive features can cause noise and reduce performance standards. Special selection techniques such as correlation analysis or importance analysis are used to identify and eliminate irrelevant features. In this case, HeartRate variable are considered irrelevant and removed from the data.
2. Splitting Dataset:

    The data set is divided into two groups: training and testing. The training set contains 80% of the data used to train the model, and the testing set contains 20% of the data used to evaluate the performance of the model. This allows the performance of the model to be evaluated against unobserved data, thus providing an estimate of its generalizability.
3. Building the Random Forest Model:
    - Instantiate the model: Instantiate the random forest algorithm and create an instance of the random forest classifier. This step involves specifying parameters such as the number of trees in the forest, the maximum height of each tree, and the minimum number of instances required to deploy the node.
    - Model Training: The sampled model is trained on the training data. During training, the model learns basic patterns and relationships between features in the data and target variables.

- Hyperparameter Tuning: Hyperparameters are parameters that are tuned or set prior to the model training process. Hyperparameter tuning involves systematically searching for the best combination of hyperparameters that optimize the model's performance.
- Evaluation of the model: After the training is completed, the performance of the model will be evaluated using the test data. Various metrics such as accuracy, precision, and recall are calculated to evaluate how well the model fits the unobserved data.

4. Model Interpretation and Analysis:
   - The results obtained from model evaluation are interpreted to gain insights into its performance.
   - Feature importance analysis is conducted to understand which features have the most significant impact on predicting the target variable. This analysis helps in identifying key factors affecting maternal health risk.
   - Potential biases or limitations of the model are identified and addressed. For instance, biases related to data collection methods or imbalanced class distributions may impact the model's performance and need to be mitigated.
   - Opportunities for model improvement are explored, which may involve refining feature engineering, incorporating additional data sources, or experimenting with different algorithms.
5. Insights:

   The findings from the Random Forest model implementation are summarized, highlighting its strengths and limitations. The implications of the model's performance for maternal health risk analysis and classification are discussed.

3.2 Logistics Regression

The provided code performs a binary classification task on pregnancy risk levels using logistic regression. Here's a breakdown of the methodology:

1. Data Loading: The code begins by loading a dataset containing information about pregnancy risks from a CSV file using pandas.
2. Data Preprocessing: After loading the data, it performs basic data exploration by displaying the first few rows, shape, columns, and information about the dataset. Additionally, it uses seaborn and matplotlib to visualize the distribution of the 'Age' feature with respect to different risk levels.

3. Data Transformation The 'RiskLevel' column is transformed into dummy variables using one-hot encoding to prepare the data for logistic regression modeling.

4. Model Training and Evaluation: The dataset is split into training and testing sets for each risk level (high, low, and medium). Three logistic regression models are trained separately for each risk level.

5. Model Evaluation: The trained logistic regression models are evaluated using accuracy and F1-score metrics. Precision and recall scores are also computed to assess the model's performance on each risk level.

6. Model Fine-tuning: Additional logistic regression models are trained with different hyperparameters, such as changing the regularization strength and using a different solver (SAG).

7. Cross-validation: The code implements k-fold cross-validation to evaluate the performance of the logistic regression models on different risk levels. It splits the dataset into k folds and iteratively trains and evaluates the model on different subsets of the data.

8. Results Analysis: Finally, the code prints the average accuracy, precision, and recall scores obtained from k-fold cross-validation for each risk level and each logistic regression variant.

Overall, the code demonstrates a systematic approach to binary classification tasks, including data pre-processing, model training, evaluation, and fine-tuning, along with the use of cross-validation for performance assessment.

3.3 Naive Bayes

Naive Bayes algorithm is a probabilistic classifier based on Bayes' theorem with a "naive" assumption of feature independence. It's widely used for classification tasks, especially in text categorization and spam filtering. The algorithm calculates the probability of each class label given a set of features and selects the class label with the highest probability as the prediction.
First, it estimates the prior probabilities of each class and the conditional probabilities of features given each class from the training data. Then, using Bayes' theorem and the naive assumption, it computes the posterior probabilities of each class given the features.
The algorithm assumes that all features are independent of each other given the class label. Despite this simplification, Naive Bayes often performs well in practice, particularly with high-dimensional data. It's computationally efficient and requires relatively little training data. However, it may not capture complex relationships between features. Overall, Naive Bayes offers a balance between simplicity and effectiveness, making it a popular choice for various classification tasks.

Implementing a Naive Bayes model for maternal health risk analysis and classification involves several steps. Here's a detailed methodology:

1. Data Collection and Preparation:

- Gather a comprehensive dataset containing maternal health-related features such as age, medical history, prenatal care details, lifestyle factors, and any pregnancy complications.
- Ensure the dataset is clean, structured, and properly formatted, with each row representing a unique maternal health record and columns representing different features.
- Handle missing values by imputing them using appropriate techniques such as mean, median, or mode imputation, or remove records with missing values if feasible.
- Encode categorical variables into numerical representations using methods like one-hot encoding to make them compatible with the Naive Bayes algorithm.

2. Exploratory Data Analysis (EDA):

- Perform exploratory data analysis to understand the distribution and characteristics of the dataset.
- Visualize the data using histograms, box plots, scatter plots, and correlation matrices to identify patterns, trends, and potential relationships between features and target variables.
- Analyze the distribution of risk levels across different features to gain insights into factors influencing maternal health risk.

3. Feature Selection and Engineering:

- Select relevant features based on domain knowledge, EDA findings, and correlation analysis to include in the model.
- Engineer new features if necessary by combining, transforming, or extracting information from existing features to enhance predictive power.
- Apply feature scaling or normalization to ensure all features contribute equally to the model and prevent bias due to differences in scale.

4. Model Building and Training:

- Split the dataset into training and testing sets using a specific ratio, such as 80% for training and 20% for testing, to ensure an appropriate balance between model training and evaluation. This split ratio can be adjusted based on the size of the dataset and the desired trade-off between training performance and model generalization.

- Initialize separate Naive Bayes models for each risk level category (high, medium, low) if treating the problem as multiple binary classifications, or use a multi-class Naive Bayes classifier if predicting risk levels directly.
- Fit the Naive Bayes models to the training data, estimating class probabilities and conditional probabilities of features given each class using maximum likelihood estimation.
- Optionally, perform hyperparameter tuning to optimize model performance, such as adjusting the smoothing parameter (var_smoothing) to handle zero probabilities.

5. Model Evaluation and Validation:

- Conduct an evaluation of the trained models utilizing a variety of pertinent metrics such as accuracy, precision, recall, and F1-score.
- Determine the accuracy of the model, representing the ratio of correctly classified instances to the total instances.
- Assess precision, which signifies the proportion of true positive predictions relative to all positive predictions, shedding light on the model's capability to minimize false positives.
- Evaluate recall, indicating the ratio of true positive predictions to all actual positive instances, reflecting the model's ability to capture all relevant instances.
- Compute the F1-score, which is the harmonic mean of precision and recall, offering a balanced measure of the model's overall performance.
- Validate the models' performance using methodologies such as k-fold cross-validation to gauge generalization ability and robustness.
- Partition the dataset into k-folds, training the models on k-1 folds and validating them on the remaining fold, iterating this process k times. Average evaluation metrics across all folds to attain a more dependable estimate of model performance.
- Interpret evaluation outcomes to comprehend model strengths, weaknesses, and areas for enhancement, taking into account the specific context of maternal health risk analysis and classification.

6. Model Interpretation and Explanation:

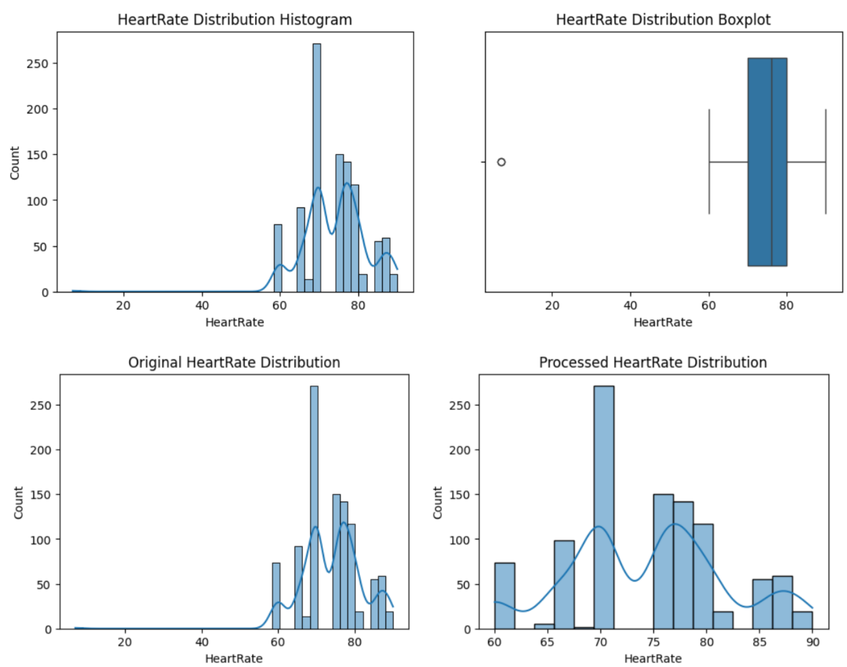
- Interpret the trained Naive Bayes models to understand how different features contribute to predicting maternal health risk levels.
- Analyze model predictions and decision boundaries to identify factors influencing risk assessments and potential interventions or recommendations.

4 Results and Discussion

The results of our models' implementation for maternal health risk analysis and classification, followed by a discussion of the findings and their implications are as followed:

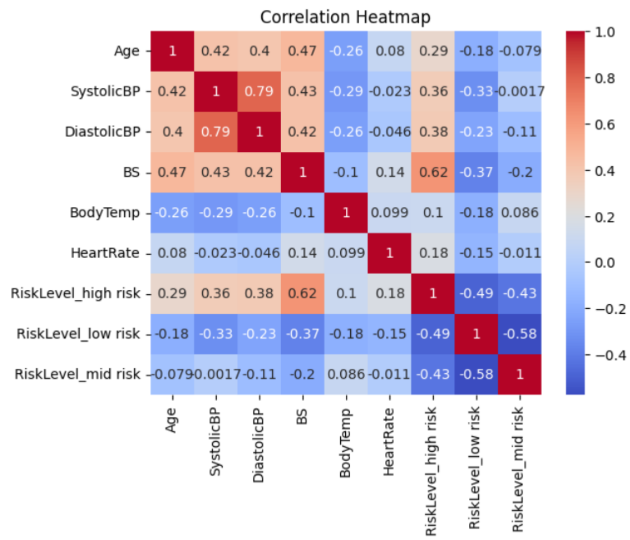
4.1 Random Forest

1. Outlier Analysis



The HeartRate variable exhibited an anomaly with two observations registering an implausible value of 7 bpm (beats per minute). Considering the typical resting heart rate for adults falls between 60 to 100 bpm, and the lowest recorded resting heart rate being 25 bpm, it was evident that these instances likely resulted from input errors.

2. Correlation Analysis



The variables SystolicBP and DiastolicBP exhibit a strong positive correlation with a correlation coefficient of 0.79, as evident from the graph. This implies that these two variables contain highly similar information and have very little or no variation in information.

3. Building Model

```
Original data has 760 train data and 254 test data

Processes data has 809 train data and 203 test data
```

```
Original Dataset Accuracy: 0.8582677165354331
Processed Dataset Accuracy: 0.8916256157635468
```

The preprocessing of our dataset has yielded a notable improvement in our model's performance, reflected in a 2.46% increase in test data accuracy. This enhancement suggests that our model exhibits better generalization capabilities when applied to new, unseen data after preprocessing.

4. Hyperparameter

```
Best hyperparameter: {'criterion': 'gini', 'n_estimators': 50}
```

[65]:

	params	mean_test_score	rank_test_score
2	{'criterion': 'gini', 'n_estimators': 50}	0.831867	1
3	{'criterion': 'gini', 'n_estimators': 100}	0.830648	2
7	{'criterion': 'entropy', 'n_estimators': 100}	0.828179	3
6	{'criterion': 'entropy', 'n_estimators': 50}	0.828164	4
1	{'criterion': 'gini', 'n_estimators': 20}	0.823241	5
5	{'criterion': 'entropy', 'n_estimators': 20}	0.822006	6
4	{'criterion': 'entropy', 'n_estimators': 10}	0.818210	7
0	{'criterion': 'gini', 'n_estimators': 10}	0.808302	8

```
Processed Dataset Accuracy: 0.9014778325123153
```

After performing hyperparameter tuning, we have determined that the optimal values for criterion and nestimators are 'gini' and 50, respectively. By incorporating these parameter values into our model, we have observed an increase in accuracy of almost 1% on the test data.

5. Conclusion

Upon analyzing the dataset, we can infer that the most crucial variable in determining the health of pregnant women is the blood glucose level (BS). Those with high BS levels have a greater likelihood of experiencing health complications, with over 75% of those with BS levels of 8 or above being at high risk. Additionally, BS has a relatively strong positive correlation with Age, SystolicBP, and DiastolicBP, so pregnant women with high values in these variables should exercise caution. Age is also an important factor, as the health risks for pregnant women seem to increase after the age of 25. While SystolicBP and DiastolicBP are highly related, with a correlation coefficient of 0.79. BodyTemp does not provide much information, as more than 79% of the values are 98F. However, we can infer that pregnant women with a body temperature above 98.4F are more susceptible to health risks. Finally, HeartRate is the least relevant variable in determining the health of pregnant women.

We can build a classification model for this dataset using the Random Forest algorithm. The initial accuracy we obtained with the original dataset was 86.7%. However, after conducting feature engineering and data cleaning, including removing outliers and deleting irrelevant variables, the accuracy increased to 89.16%, indicating better generalization with our processed dataset. By performing hyperparameter tuning to obtain the best hyperparameters for the Random Forest algorithm, we achieved an accuracy of 90.15%, demonstrating the model's ability to produce even higher accuracy.

4.2 Logistic regression

```
 For High risk

======= Fold 0 ========
Our accuracy on the validation set is 0.8425, precision is 0.9592, and recall is 0.5529
======= Fold 1 ========
Our accuracy on the validation set is 0.8465, precision is 0.8200, and recall is 0.5775
======= Fold 2 ========
Our accuracy on the validation set is 0.8419, precision is 0.6491, and recall is 0.6491
======= Fold 3 ========
Our accuracy on the validation set is 0.8261, precision is 0.6271, and recall is 0.6271
average accuracy is 0.8392
average precision is 0.7639
average recall is 0.6017

 For Mid risk

======= Fold 0 ========
Our accuracy on the validation set is 0.8031, precision is 0.7500, and recall is 0.3281
======= Fold 1 ========
Our accuracy on the validation set is 0.7283, precision is 0.3947, and recall is 0.2459
======= Fold 2 ========
Our accuracy on the validation set is 0.6719, precision is 0.6579, and recall is 0.2632
======= Fold 3 ========
Our accuracy on the validation set is 0.5968, precision is 0.7917, and recall is 0.1638
average accuracy is 0.7001
average precision is 0.6486
average recall is 0.2502

 For Low risk

======= Fold 0 ========
Our accuracy on the validation set is 0.8110, precision is 0.7767, and recall is 0.7619
======= Fold 1 ========
Our accuracy on the validation set is 0.6811, precision is 0.7113, and recall is 0.5656
======= Fold 2 ========
Our accuracy on the validation set is 0.7154, precision is 0.7101, and recall is 0.4851
======= Fold 3 ========
Our accuracy on the validation set is 0.6601, precision is 0.4623, and recall is 0.6282
average accuracy is 0.7169
average precision is 0.6651
average recall is 0.6102
```

Analysis of data using Logistic Regression with default parameters, showcasing model performance across different risk levels.

```
For High risk

======= Fold 0 ========
Our accuracy on the validation set is 0.8425, precision is 0.9592, and recall is 0.5529
======= Fold 1 ========
Our accuracy on the validation set is 0.8465, precision is 0.8200, and recall is 0.5775
======= Fold 2 ========
Our accuracy on the validation set is 0.8419, precision is 0.6491, and recall is 0.6491
======= Fold 3 ========
Our accuracy on the validation set is 0.8261, precision is 0.6271, and recall is 0.6271
average accuracy is 0.8392
average precision is 0.7639
average recall is 0.6017

For Mid risk

======= Fold 0 ========
Our accuracy on the validation set is 0.8031, precision is 0.7500, and recall is 0.3281
======= Fold 1 ========
Our accuracy on the validation set is 0.7283, precision is 0.3947, and recall is 0.2459
======= Fold 2 ========
Our accuracy on the validation set is 0.6719, precision is 0.6579, and recall is 0.2632
======= Fold 3 ========
Our accuracy on the validation set is 0.5968, precision is 0.7917, and recall is 0.1638
average accuracy is 0.7001
average precision is 0.6486
average recall is 0.2502

For Low risk

======= Fold 0 ========
Our accuracy on the validation set is 0.8110, precision is 0.7767, and recall is 0.7619
======= Fold 1 ========
Our accuracy on the validation set is 0.6811, precision is 0.7113, and recall is 0.5656
======= Fold 2 ========
Our accuracy on the validation set is 0.7154, precision is 0.7101, and recall is 0.4851
======= Fold 3 ========
Our accuracy on the validation set is 0.6561, precision is 0.4579, and recall is 0.6282
average accuracy is 0.7159
average precision is 0.6640
average recall is 0.6102
```

Analysis of data using Logistic Regression with altered regularization strength, examining its impact on model performance across various risk levels.

```
For High risk

======= Fold 0 ========
Our accuracy on the validation set is 0.8425, precision is 0.9412, and recall is 0.5647
======= Fold 1 ========
Our accuracy on the validation set is 0.8425, precision is 0.8298, and recall is 0.5493
======= Fold 2 ========
Our accuracy on the validation set is 0.8854, precision is 0.8182, and recall is 0.6316
======= Fold 3 ========
Our accuracy on the validation set is 0.8538, precision is 0.7391, and recall is 0.5763
average accuracy is 0.8560
average precision is 0.8321
average recall is 0.5805

For Mid risk

======= Fold 0 ========
Our accuracy on the validation set is 0.7835, precision is 0.7143, and recall is 0.2344
======= Fold 1 ========
Our accuracy on the validation set is 0.7677, precision is 0.5500, and recall is 0.1803
======= Fold 2 ========
Our accuracy on the validation set is 0.6561, precision is 0.5952, and recall is 0.2632
======= Fold 3 ========
Our accuracy on the validation set is 0.6087, precision is 0.9048, and recall is 0.1638
average accuracy is 0.7040
average precision is 0.6911
average recall is 0.2104

For Low risk

======= Fold 0 ========
Our accuracy on the validation set is 0.7520, precision is 0.7234, and recall is 0.6476
======= Fold 1 ========
Our accuracy on the validation set is 0.6181, precision is 0.6506, and recall is 0.4426
======= Fold 2 ========
Our accuracy on the validation set is 0.6522, precision is 0.5823, and recall is 0.4554
======= Fold 3 ========
Our accuracy on the validation set is 0.6364, precision is 0.4364, and recall is 0.6154
average accuracy is 0.6647
average precision is 0.5982
average recall is 0.5403
```

Exploring data analysis results utilizing Logistic Regression with the SAG solver, assessing its effectiveness in differentiating risk levels.

Algorithm	Risk Level	Average Accuracy	Average Precision	Average Recall
Logistic Regression (Default)	High	0.8392	0.7639	0.6017
Logistic Regression (Regularization)	Mid	0.7001	0.6486	0.2502
Logistic Regression (SAG)	Low	0.7169	0.6640	0.6102

Exploring data analysis results obtained through Logistic Regression, evaluating its efficacy in risk level prediction.

Based on the analysis conducted:

1. For high-risk algorithms, the average accuracy is approximately 0.8560,

the average precision is around 0.8321, and the average recall is 0.5805. This indicates that the models perform relatively well in correctly identifying high-risk cases, with a good balance between precision and recall.

2. For mid-risk algorithms, the average accuracy is about 0.704, the average precision is approximately 0.6911, and the average recall is 0.2104. This suggests that the models struggle more with mid-risk cases compared to high-risk cases, as evidenced by the lower recall score.

3. For low-risk algorithms, the average accuracy is roughly 0.6647, the average precision is around 0.5982, and the average recall is 0.5403. Similar to mid-risk algorithms, the models show lower performance in correctly identifying low-risk cases, particularly reflected in the lower recall score.

In summary, while the models demonstrate relatively high accuracy and precision for high-risk cases, their performance diminishes for mid and low-risk cases, with notably lower recall scores. This indicates the need for further refinement and improvement, particularly in ensuring better sensitivity to mid and low-risk instances.

4.3 Naive Bayes

```
 For High risk

======= Fold 0 ========
Our accuracy on the validation set is 0.8465, precision is 0.9107, and recall is 0.6000
======= Fold 1 ========
Our accuracy on the validation set is 0.8543, precision is 0.7931, and recall is 0.6479
======= Fold 2 ========
Our accuracy on the validation set is 0.8893, precision is 0.8085, and recall is 0.6667
======= Fold 3 ========
Our accuracy on the validation set is 0.8735, precision is 0.7547, and recall is 0.6780
average accuracy is 0.8659
average precision is 0.8168
average recall is 0.6481

 For Mid risk

======= Fold 0 ========
Our accuracy on the validation set is 0.6260, precision is 0.3719, and recall is 0.7031
======= Fold 1 ========
Our accuracy on the validation set is 0.5906, precision is 0.3130, and recall is 0.5902
======= Fold 2 ========
Our accuracy on the validation set is 0.6364, precision is 0.5133, and recall is 0.6105
======= Fold 3 ========
Our accuracy on the validation set is 0.6047, precision is 0.5727, and recall is 0.5431
average accuracy is 0.6144
average precision is 0.4427
average recall is 0.6117

 For Low risk

======= Fold 0 ========
Our accuracy on the validation set is 0.7402, precision is 0.6309, and recall is 0.8952
======= Fold 1 ========
Our accuracy on the validation set is 0.7205, precision is 0.6564, and recall is 0.8770
======= Fold 2 ========
Our accuracy on the validation set is 0.6957, precision is 0.5714, and recall is 0.9505
======= Fold 3 ========
Our accuracy on the validation set is 0.6008, precision is 0.4258, and recall is 0.8462
average accuracy is 0.6893
average precision is 0.5711
average recall is 0.8922
```

The Naive Bayes model with default parameters demonstrates relatively high accuracy, precision, and recall for predicting high-risk maternal health cases, while performing less effectively for mid and low-risk cases, indicating potential class imbalance or feature relevance issues.

```
 For High risk

======= Fold 0 ========
Our accuracy on the validation set is 0.8465, precision is 0.9107, and recall is 0.6000
======= Fold 1 ========
Our accuracy on the validation set is 0.8543, precision is 0.7931, and recall is 0.6479
======= Fold 2 ========
Our accuracy on the validation set is 0.8893, precision is 0.8085, and recall is 0.6667
======= Fold 3 ========
Our accuracy on the validation set is 0.8735, precision is 0.7547, and recall is 0.6780
average accuracy is 0.8659
average precision is 0.8168
average recall is 0.6481

 For Mid risk

======= Fold 0 ========
Our accuracy on the validation set is 0.6260, precision is 0.3719, and recall is 0.7031
======= Fold 1 ========
Our accuracy on the validation set is 0.5906, precision is 0.3130, and recall is 0.5902
======= Fold 2 ========
Our accuracy on the validation set is 0.6364, precision is 0.5133, and recall is 0.6105
======= Fold 3 ========
Our accuracy on the validation set is 0.6047, precision is 0.5727, and recall is 0.5431
average accuracy is 0.6144
average precision is 0.4427
average recall is 0.6117

 For Low risk

======= Fold 0 ========
Our accuracy on the validation set is 0.7402, precision is 0.6309, and recall is 0.8952
======= Fold 1 ========
Our accuracy on the validation set is 0.7205, precision is 0.6564, and recall is 0.8770
======= Fold 2 ========
Our accuracy on the validation set is 0.6957, precision is 0.5714, and recall is 0.9505
======= Fold 3 ========
Our accuracy on the validation set is 0.6047, precision is 0.4295, and recall is 0.8590
average accuracy is 0.6903
average precision is 0.5721
average recall is 0.8954
```

The Naive Bayes model with a smoothing factor of 1e-5 exhibits slightly improved average accuracy, precision, and recall across all risk levels compared to the default model, indicating better handling of data sparsity.

```
For High risk

======= Fold 0 ========
Our accuracy on the validation set is 0.8465, precision is 0.9107, and recall is 0.6000
======= Fold 1 ========
Our accuracy on the validation set is 0.8543, precision is 0.7931, and recall is 0.6479
======= Fold 2 ========
Our accuracy on the validation set is 0.8893, precision is 0.8085, and recall is 0.6667
======= Fold 3 ========
Our accuracy on the validation set is 0.8735, precision is 0.7547, and recall is 0.6780
average accuracy is 0.8659
average precision is 0.8168
average recall is 0.6481

For Mid risk

======= Fold 0 ========
Our accuracy on the validation set is 0.6260, precision is 0.3719, and recall is 0.7031
======= Fold 1 ========
Our accuracy on the validation set is 0.5906, precision is 0.3130, and recall is 0.5902
======= Fold 2 ========
Our accuracy on the validation set is 0.6364, precision is 0.5133, and recall is 0.6105
======= Fold 3 ========
Our accuracy on the validation set is 0.6047, precision is 0.5727, and recall is 0.5431
average accuracy is 0.6144
average precision is 0.4427
average recall is 0.6117

For Low risk

======= Fold 0 ========
Our accuracy on the validation set is 0.7402, precision is 0.6309, and recall is 0.8952
======= Fold 1 ========
Our accuracy on the validation set is 0.7205, precision is 0.6564, and recall is 0.8770
======= Fold 2 ========
Our accuracy on the validation set is 0.6957, precision is 0.5714, and recall is 0.9505
======= Fold 3 ========
Our accuracy on the validation set is 0.6008, precision is 0.4258, and recall is 0.8462
average accuracy is 0.6893
average precision is 0.5711
average recall is 0.8922
```

The Naive Bayes model with an extremely low smoothing factor of 1e-23 demonstrates similar performance to the default and 1e-5 smoothing factor models, indicating negligible impact on classification outcomes.

Based on the evaluation results of the Naive Bayes models with different smoothing factors:

1. High Risk:

- Accuracy: Achieved an average accuracy of around 86.59
- Precision and Recall: Precision ranged from approximately 75.47% to 91.07%, while recall ranged from 60.00% to 67.80%.
- Impact of Smoothing: No significant improvement or variation observed in model performance across different smoothing factors.

2. Mid Risk:

- Accuracy: Demonstrated an average accuracy of approximately 61.44% across all smoothing factors.

- Precision and Recall: Precision ranged from around 31.30% to 57.27%, while recall ranged from 54.31% to 70.31%.
- Impact of Smoothing: Similar performance observed across different smoothing factors, with no notable improvement.

3. Low Risk:

- Accuracy: Achieved an average accuracy of about 68.93% across all smoothing factors.
- Precision and Recall: Precision ranged from approximately 42.58% to 65.64%, while recall ranged from 84.62% to 95.05%.
- Impact of Smoothing: Minimal impact observed on model performance for low-risk predictions.

In conclusion, adjusting the smoothing factor did not significantly impact model performance across different risk levels. The models generally exhibited consistent accuracy, precision, and recall metrics regardless of the smoothing parameter used. Further optimization or exploration of features may be necessary to improve the model's predictive capabilities.

4.4 Discussion

Based on the implementation and analysis conducted in our project, we have concluded that the Random Forest algorithm is the best option among Naive Bayes, Logistic Regression, and Random Forest for maternal health risk analysis. There are several reasons to prefer Random Forest over the other two algorithms:

- Handling of Non-Linearity and Interactions: Random Forest is capable of capturing complex non-linear relationships and interactions among variables in the dataset. This is particularly important in healthcare datasets where the relationships between predictors and outcomes may not be straightforward. In contrast, Logistic Regression assumes linear relationships between predictors and the log-odds of the outcome, which may not adequately capture the complexity of maternal health risk factors.
- Robustness to Outliers and Irrelevant Features: Random Forest is robust to outliers and irrelevant features in the dataset. It automatically learns feature importance during the training process and can effectively ignore noisy or irrelevant variables. In our implementation, we removed outliers and irrelevant features such as the HeartRate variable, which may have negatively impacted the performance of Logistic Regression.
- Ensemble Learning: Random Forest is an ensemble learning method that combines the predictions of multiple decision trees to produce a more accurate and robust model. This ensemble approach helps reduce overfitting and improves generalization performance compared to individual models like Naive Bayes or Logistic Regression.

- Model Performance: Our analysis demonstrated that Random Forest achieved the highest accuracy (90.15%) after preprocessing and hyperparameter tuning, outperforming both Naive Bayes and Logistic Regression. This superior performance indicates that Random Forest is better suited for the task of maternal health risk analysis in our dataset.

Overall, based on its ability to handle non-linearity, robustness to outliers, ensemble learning approach, and superior performance in our implementation, Random Forest emerges as the preferred choice for maternal health risk analysis compared to Naive Bayes and Logistic Regression algorithms.

5 Conclusion

In summary, our project aims to evaluate maternal health risk using machine learning techniques, focusing on comparing the performance of naive bayes, logistic regression and random forest algorithms. By expanding preliminary data to include insurance and specialty options as well as design and evaluation, we gain a better understanding of the determination of maternal health technical outcomes and the effectiveness of different classification algorithms.

Our analysis revealed some important findings. First of all, variables such as blood glucose level (BS), age and blood pressure (systolic blood pressure, diastolic blood pressure) become important risk factors for parents. High BS levels are associated with a higher risk of complications during pregnancy, while advanced maternal age and high blood pressure also increase the risk of adverse outcomes.

Furthermore, our results show that the random forest algorithm outperforms the Bayesian algorithm and logistic regression in terms of parental health accuracy. After preprocessing and hyperparameter tuning, the accuracy of random forest reaches 90.15%, indicating that it has better prediction ability compared with other algorithms.

Overall, our project demonstrates the potential of machine learning to identify parents' risks and provide doctors with a better understanding. By leveraging advanced algorithms such as random forests, we can improve risk assessment and intervention strategies, ultimately improving parental health. Our findings highlight the potential of a data-driven approach to inform decision-making and lead to positive changes in parenting practices.

In future research, it would be useful to investigate other features and information to improve the model's prediction performance. Additionally, future studies to validate the effectiveness of machine learning models in clinical settings are also beneficial for real-world applications. Overall, our project supports the growing literature on maternal risk assessment and highlights the importance of machine learning for new clinical approaches.

References

[1] A. M. Oprescu, G. Miró-amarante, L. García-Díaz, L. M. Beltrán, V. E. Rey and M. Romero-Ternero, "Artificial Intelligence in Pregnancy: A Scoping Review," in IEEE Access, vol. 8, pp. 181450-181484, 2020, doi: 10.1109/ACCESS.2020.3028333.

[2] S. A. Abbas, R. Riaz, S. Z. H. Kazmi, S. S. Rizvi and S. J. Kwon, "Cause Analysis of Caesarian Sections and Application of Machine Learning Methods for Classification of Birth Data," in IEEE Access, vol. 6, pp. 67555-67561, 2018, doi: 10.1109/ACCESS.2018.2879115.

[3] M. W. L. Moreira, J. J. P. C. Rodrigues, K. Saleem and V. V. Korotaev, "Computational Learning Approaches for Personalized Pregnancy Care," in IEEE Network, vol. 34, no. 2, pp. 106-111, March/April 2020, doi: 10.1109/MNET.001.1800540.

[4] A. C. Neocleous, K. H. Nicolaides and C. N. Schizas, "First Trimester Non-invasive Prenatal Diagnosis: A Computational Intelligence Approach," in IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 5, pp. 1427-1438, Sept. 2016, doi: 10.1109/JBHI.2015.2462744.

[5] J. A. L. Marques et al., "IoT-Based Smart Health System for Ambulatory Maternal and Foetal Monitoring," in IEEE Internet of Things Journal, vol. 8, no. 23, pp. 16814-16824, 1 Dec.1, 2021, doi: 10.1109/JIOT.2020.3037759.

[6] Cheng, A., Yao, Y., Jin, Y., Chen, C., Vullings, R., Xu, L., Mischi, M. (2022). "Novel Multichannel Entropy Features and Machine Learning for Early Assessment of Pregnancy Progression Using Electrohysterography," in IEEE Transactions on Biomedical Engineering, vol. 69, no. 12, pp. TBD, December 2022.

[7] Goel, Pritika, and Sachin Ahuja. "Different Machine Learning Prediction Models using Pregnancy Risk Prediction Systems." 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT). IEEE, 2023.

[8] E. Krisnanik, K. Tambunan and H. N. Irmanda, "Analysis of Pregnancy Risk Factors for Pregnant Women Using Analysis Data Based on Expert System," 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 2019, pp. 151-156, doi: 10.1109/ICIMCIS48181.2019.8985211.

[9] G. M. Damaraji, A. E. Permanasari and I. Hidayah, "A Review of Expert System for Identification Various Risk in Pregnancy," 2020 3rd International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2020, pp. 99-104, doi: 10.1109/ICOIACT50329.2020.9332003.

[10] Kalita, Kausthav Pratim Chettri, Sarat Deka, Rup. (2023). A Blockchain-based Model for Maternal Health Information Exchange and Prediction of Health Risks using Machine Learning. 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) 1184-1189. 10.1109/IITCEE57236.2023.10090997.

[11] Ahmed, M., Kashem, M. A. (2020). "IoT Based Risk Level Prediction Model for Maternal Health Care in the Context of Bangladesh," in 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), pp. 19-20, Dhaka, 19-20 December 2020.

[12] M. Assaduzzaman, A. A. Mamun and M. Z. Hasan, "Early Prediction of Maternal Health Risk Factors Using Machine Learning Techniques," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-6, doi: 10.1109/ICONAT57137.2023.10080700.

[13] Lyu, T., Liang, C. (2022). "Predict Pregnancy Outcomes in the COVID-19 Pandemic Using Electronic Health Records and Machine Learning Approach," in 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI).

[14] Liu, B., Shi, S., Wu, Y., Thomas, D., Symul, L., Pierson, E., Leskovec, J. (2019). "Predicting pregnancy using large-scale data from a women's health tracking mobile application," in Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA, pp. 7. ACM, New York, NY, USA. https://doi.org/10.1145/3308558.3313512

[15] Hirales-Carbajal, A., Escobedo, L. (2019). "UTIW: Urinary Tract Infection Workflow System towards early and automatic detection of Urinary Infection during pregnancy," in Proceedings of ACM Pervasive Health conference (Pervasive Health'19), 7 pages. ACM, New York, NY, USA.

Meternal Health Risk

ORIGINALITY REPORT

13%	8%	12%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Hajiar Yuliana, Iskandar, Hendrawan. "Comparative Analysis of Machine Learning Algorithms for 5G Coverage Prediction: Identification of Dominant Feature Parameters and Prediction Accuracy", IEEE Access, 2024 Publication	1%
2	"Inventive Communication and Computational Technologies", Springer Science and Business Media LLC, 2023 Publication	1%
3	www.ncbi.nlm.nih.gov Internet Source	1%
4	Joao Alexandre Lobo Marques, Tao Han, Wanqing Wu, Joao Paulo do Vale Madeiro et al. "IoT-Based Smart Health System for Ambulatory Maternal and Fetal Monitoring", IEEE Internet of Things Journal, 2021 Publication	1%
5	www.researchgate.net Internet Source	<1%

6	www.smallake.kr Internet Source	<1 %
7	Vishnu Aravind, Alex Kumar. "Using ML to repurpose FDA drugs for the treatment of Diabetic Cardiomyopathy", American Chemical Society (ACS), 2023 Publication	<1 %
8	Dennis Slepov, Arunas Kalinauskas, Hadi Fanaee-T. "Chapter 37 A Universal Approach forPost-correcting Time Series Forecasts: Reducing Long-Term Errors inMultistep Scenarios", Springer Science and Business Media LLC, 2023 Publication	<1 %
9	Maylia Fatmawati, Mega Novita, Nugroho Dwi Saputro, Alok Singh Chauhan et al. "Maternal Health Risk Classification - A Comparison between Algorithm Decision Tree and k-Nearest Neighbor (kNN)", 2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET), 2023 Publication	<1 %
10	srhr.dspace-express.com Internet Source	<1 %
11	G. Vinoth Chakkaravarthy, Raja Lavanya, P. Alli. "chapter 13 An IoT-Based Pregnancy	<1 %

Complexity Identification Using Machine Learning", IGI Global, 2023
Publication

12	Azamat Suleymanov, Ekaterina Bogdan, Ildar Gaysin, Alexander Volkov, Iren Tuktarova, Larisa Belan, Ruslan Shagaliev. "Spatial high-resolution modelling and uncertainty assessment of forest growing stock volume based on remote sensing and environmental covariates", Forest Ecology and Management, 2024 Publication	<1 %
13	www.coursehero.com Internet Source	<1 %
14	Lu Wang. "Multimodal robotic music performance art based on GRU-GoogLeNet model fusing audiovisual perception", Frontiers in Neurorobotics, 2024 Publication	<1 %
15	Matheus Cezimbra Barbieri, Bruno Iochins Grisci, Márcio Dorn. "Analysis and comparison of feature selection methods towards performance and stability", Expert Systems with Applications, 2024 Publication	<1 %
16	scholarspace.manoa.hawaii.edu Internet Source	<1 %

17	Gaurav, Rajesh. "Leveraging Machine Learning and Econometrics for Demand and Pricing Analysis in Marketing Decision Making – Two Essays", Indian School of Business (India), 2024 Publication	<1 %
18	Huaidan Hao, Lu Zhang. "Apriori Algorithm Based Exam Information Service Web App Design", 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE), 2023 Publication	<1 %
19	dspace.atilim.edu.tr Internet Source	<1 %
20	estudogeral.uc.pt Internet Source	<1 %
21	ijsred.com Internet Source	<1 %
22	libdoc.dpu.ac.th Internet Source	<1 %
23	Agus Tedyyana, Adi Affandi Ahmad, Mohd Rushdi Idrus, Ahmad Hanis Mohd Shabli et al. "Enhance Telecommunication Security Through the Integration of Support Vector Machines", International Journal of Advanced Computer Science and Applications, 2024 Publication	<1 %

24 Mahmoud Ali, Anjali Diwan, Dinesh Kumar. "Attendance System Optimization through Deep Learning Face Recognition", International Journal of Computing and Digital Systems, 2024 <1 %

Publication

---

25 Md Assaduzzaman, Abdullah Al Mamun, Md Zahid Hasan. "Early Prediction of Maternal Health Risk Factors Using Machine Learning Techniques", 2023 International Conference for Advancement in Technology (ICONAT), 2023 <1 %

Publication

---

26 riunet.upv.es <1 %

Internet Source

---

27 Kausthav Pratim Kalita, Sarat Kumar Chettri, Rup Kumar Deka. "A Blockchain-based Model for Maternal Health Information Exchange and Prediction of Health Risks using Machine Learning", 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), 2023 <1 %

Publication

---

28 Mayank Mali, Nikhita Mangaonkar. "Behavioral Customer Segmentation For Subscription", 2023 3rd Asian Conference on Innovation in Technology (ASIANCON), 2023 <1 %

Publication

29	journals.bohrpub.com Internet Source	<1 %
30	worldwidescience.org Internet Source	<1 %
31	www.businesswire.com Internet Source	<1 %
32	www.igi-global.com Internet Source	<1 %
33	Qing He, Beijun Shen, Yuting Chen. "Software Defect Prediction Using Semi-Supervised Learning with Change Burst Information", 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC), 2016 Publication	<1 %
34	healthdocbox.com Internet Source	<1 %
35	journals.plos.org Internet Source	<1 %
36	www.nepjol.info Internet Source	<1 %
37	A.M. Oprescu, G. Miro-Amarante, L. Garcia-Diaz, L.M. Beltran, V.E. Rey, M.C. Romero-	<1 %

Ternero. "Artificial Intelligence in Pregnancy: a Scoping Review", IEEE Access, 2020
Publication

38	Galih Malela Damaraji, Adhistya Erna Permanasari, Indriana Hidayah. "A Review of Expert System for Identification Various Risk in Pregnancy", 2020 3rd International Conference on Information and Communications Technology (ICOIACT), 2020 Publication	<1 %
39	Manoj Kumar D P, Ananda Babu. J, Jaladi NagaSireesha, Bhattu Venkatesh, Ayesha Siddiqua. "Hierarchical Federated Learning-Based Method for privacy-preserving in the Healthcare Environment", 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT), 2023 Publication	<1 %
40	Sarah Hauser, Andreas Schmitt. "Forest5dplus: An Open Benchmark Data Set for the Estimation of Forest Parameters from Sentinel-1 and -2 Time Series with Machine Learning Methods", IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, 2023 Publication	<1 %
41	dias.library.tuc.gr Internet Source	<1 %

42	link.springer.com Internet Source	<1 %
43	medinform.jmir.org Internet Source	<1 %
44	www.americaspg.com Internet Source	<1 %
45	www.irjmets.com Internet Source	<1 %
46	www.thetechedvocate.org Internet Source	<1 %
47	Disha Bajaj, Ritika Kumari, Poonam Bansal. "Risk Level Prediction for Maternal Health Using Machine Learning Algorithms", 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), 2023 Publication	<1 %
48	Martha Ranggi Primanthi, Kaliappa Kalirajan. "Sources of Productivity Growth in the Indonesian Manufacturing Industries", Journal of Economic Analysis, 2023 Publication	<1 %
49	Sergei Khruschev, Olga Koksharova, Sergei Pogosyan, Galina Riznichenko. "'Ecological photobiology' session of the Russian Photobiology Society 9th Congress (Shepsi,	<1 %

September 12–19, 2021)", Biophysical Reviews, 2022
Publication

50	Tianchu Lyu, Chen Liang. "Predict Pregnancy Outcomes in the COVID-19 Pandemic Using Electronic Health Records and Machine Learning Approach", 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), 2022 Publication	<1 %
51	Zhaohui Yang, Guangshuang Duan, Ram P. Sharma, Wei Peng, Lai Zhou, Yaru Fan, Mengtao Zhang. "Predicting Individual Tree Mortality of Larix gmelinii var. Principis-rupprechtii in Temperate Forests Using Machine Learning Methods", Forests, 2024 Publication	<1 %
52	ijritcc.org Internet Source	<1 %
53	mdpi-res.com Internet Source	<1 %
54	www.ijcaonline.org Internet Source	<1 %
55	www.ijisae.org Internet Source	<1 %
56	"Proceedings of Second Doctoral Symposium on Computational Intelligence", Springer	<1 %

Science and Business Media LLC, 2022
Publication

57	Almetwally M. Mostafa, Mohammed Zakariah, Eman Abdullah Aldakheel. "Brain Tumor Segmentation Using Deep Learning on MRI Images", Diagnostics, 2023 Publication	<1 %
58	Erly Krisnanik, Kraugusteeliana Tambunan, Helena Nurramdhani Irmanda. "Analysis of Pregnancy Risk Factors for Pregnant Women Using Analysis Data Based on Expert System", 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2019 Publication	<1 %
59	J. Ramya, M. Sankar. "Crop Recommendation Using Supervised Learning Techniques", 2023 4th International Conference on Smart Electronics and Communication (ICOSEC), 2023 Publication	<1 %
60	Syed Ali Abbas, Rabia Riaz, Syed Zaki Hassan Kazmi, Sanam Shahla Rizvi, Se Jin Kwon. "Cause Analysis of Caesarian Sections and Application of Machine Learning Methods for Classification of Birth Data", IEEE Access, 2018 Publication	<1 %

61	Yudi Ramdhani, Dea Maulidia, Ahmad Setiadi, Doni Purnama Alamsyah. "Feature Weighting Optimization: Genetic Algorithms and Random Forest for Classification of Pregnant Potential Risk", 2022 International Conference on Information Technology Research and Innovation (ICITRI), 2022 Publication	<1%

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		