CIS530 Final Project: Predictive Modeling of Maternal Health Risk Factors Using Machine Learning Techniques

Preetham Thatikonda
DataScience
Umass,Dartmouth
Dartmouth, Massachusetts
pthatikonda@umassd.edu
02182487

Pavan Kumar Etta
Data science
Umass Dartmouth
Dartmouth, Massachusetts
petta@umassd.edu
02179303

Sarvagna Kotte
Data Science
Umass,Dartmouth
Dartmouth, Massachusetts
skotte@umassd.edu
02185847

Roshan Pampari
DataScience
Umass,Dartmouth
Dartmouth, Massachusetts
rpampari@umassd.edu
02143843

Yogi Kodali
Data science
Umass Dartmouth
Dartmouth, Massachusetts
ykodali@umassd.edu
02168810

***Abstract*— This study focuses on predicting maternal health risk levels (high, mid, low) using clinical features such as blood pressure, blood sugar, body temperature, and heart rate. We applied Random Forest and Logistic Regression models, evaluated via accuracy, confusion matrices, and ROC curves. Using 5-fold cross-validation and both 50/50 and 80/20 train-test splits, Random Forest achieved the highest accuracy of 84.65%. While effective, limitations include dataset size and class imbalance, indicating a need for broader validation.**

Keywords—Maternal Health, Risk Prediction, Machine Learning, Random Forest, Logistic Regression, Data Mining, Healthcare Analytics

I. INTRODUCTION

Maternal health is a critical aspect of public health, particularly in the post-delivery period when women are vulnerable to various complications. Timely identification of individuals at high risk can significantly improve health outcomes through early interventions and prioritized care [6]. Predictive modeling using clinical data offers a promising avenue to support healthcare providers in making informed decisions [1][7]. In this study, we aim to develop a reliable machine learning model to predict maternal health risk levels—categorized as high, mid, or low—based on physiological indicators such as blood pressure, blood sugar, body temperature, and heart rate. Using a publicly available dataset with 1,014 instances and no missing values, we applied data mining techniques, focusing on interpretable models to ensure clinical applicability [2][3]. Our goal is to contribute to the integration of predictive analytics into maternal healthcare, ultimately supporting better resource allocation and enhancing patient outcomes [1][4].

II. DATASET

A. Dataset Description

The dataset used in this study was sourced from a publicly available maternal health risk dataset, such as those hosted on platforms like the UCI Machine Learning Repository. It consists of 1,014 clinical records, each capturing important physiological indicators relevant to maternal health [2]. The dataset includes attributes such as age, systolic blood pressure, diastolic blood pressure, blood sugar levels, body temperature, and heart rate, with the target variable being RiskLevel categorized as High, Mid, or Low. All records were verified for clinical relevance, ensuring that there were no missing values, which enhances the robustness and reliability of the subsequent data mining analysis [3].

B. Data Integrity and Preprocessing

Prior to analysis, the dataset underwent rigorous preprocessing to ensure consistency and readiness for machine learning tasks [1]. No null or missing values were found, and all features were assessed to confirm they fell within expected clinical ranges (e.g., systolic blood pressure between 90–180 mmHg, heart rate between 60–100 bpm) [5]. The target variable RiskLevel was encoded as a categorical factor with three classes. Numerical attributes were normalized to maintain model compatibility and improve training efficiency. These steps helped preserve the integrity of the data while enhancing the robustness of subsequent analytical models.

III. LITERATURE SURVEY

Predicting maternal health has emerged as an important area in research, driven by the need to reduce maternal mortality and improve maternal outcomes. Various machine learning (ML) methods have been used in recent studies to identify high-risk pregnancies and offer early interventions.

A recent study by Al Hinai et al. evaluated ten different ML methods, including Logistic Regression, Random Forest, and Decision Trees, on a dataset of 402 maternal deaths in Oman spanning from 1991 to 2023. The dataset was preprocessed using Principal Component Analysis (PCA) to reduce dimensionality. Among the models tested, Random Forest achieved the highest accuracy of 75.2%, highlighting the effectiveness of classification methods like Random Forest in the health domain [8].

In another study, Oluwarotimi et al. proposed a deep hybrid model combining Artificial Neural Networks (ANN) with Random Forest for maternal risk classification. Using clinical parameters such as age, blood pressure, and heart rate, the model achieved an impressive accuracy of 95%. The synergy of ANN's deep feature extraction with Random Forest's robustness demonstrates the potential of hybrid approaches in health risk prediction [9].

Pratama et al. conducted an evaluation of multiple ML models for predicting maternal health risk levels and emphasized model benchmarking in clinical datasets [10].

Patel et al. employed Explainable Boosting Machines (EBMs) to predict severe maternal morbidity and preeclampsia. EBMs offered higher accuracy than traditional logistic regression models while maintaining interpretability—a critical factor in clinical decision-making [11].

Lastly, Park et al. conducted a comparative study between ML algorithms and logistic regression for predicting adverse outcomes in preeclampsia. While ML models such as Random Forest provided better discrimination, logistic regression demonstrated superior calibration, highlighting the trade-offs between model performance and reliability in medical settings [12].

IV. METHODOLOGY

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

A. Data Cleaning and Validation

The dataset comprised 1,014 instances with no missing values. A thorough check was conducted to verify the clinical relevance of feature ranges, such as ensuring blood pressure values lay between 90–180 mmHg and heart rates ranged from 60–100 bpm. This validation ensured that the input data reflected real-world physiological conditions, reinforcing the dataset's reliability [4].

B. Exploratory Data Analysis

Exploratory analysis was performed to better understand feature distributions and relationships. Bar plots were used to visualize individual attributes, while a correlation heatmap highlighted the interactions between variables. Notably, a strong correlation was observed between systolic and diastolic blood pressure, indicating potential multicollinearity [5].

C. Feature Transformation

To prepare for classification, the target variable RiskLevel was encoded as a factor with three levels: High, Mid, and Low. All numerical features were normalized to standardize scales across attributes, enhancing model training and comparability across different algorithms [2].

D. Modeling Approach

In this study, two classification models—Random Forest and Logistic Regression—were employed to predict maternal health risk levels (High, Mid, Low) based on physiological indicators. The choice of models was driven by the need for both predictive performance and interpretability [1][4]. To ensure reliability, each model underwent thorough evaluation using standard metrics and a robust cross-validation approach.

Models Implemented :

Random Forest: This ensemble-based method was chosen for its ability to handle overfitting and noise in tabular clinical data. It also offers high accuracy and interpretability through feature importance measures [4][5].

Logistic Regression: Selected as a baseline model, it is known for its simplicity and interpretability. It efficiently handles multiclass classification problems using a one-vs-rest (OvR) strategy [3].

To assess model performance, various classification metrics were used, including accuracy, confusion matrices, and ROC curves. These helped evaluate how well each model could distinguish between the High, Mid, and Low risk categories. The following formula represents the basic structure of logistic regression used for prediction:

$$P(Y = k|X) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kn}x_n}}{\sum_{j=1}^{K} e^{\beta_{j0}+\beta_{j1}x_1+\cdots+\beta_{jn}x_n}} \quad (1)$$

Equation (1) illustrates the softmax function used in multinomial logistic regression, where $P(Y=k|X)$ is the probability of class k given input features X, and β represents model coefficients [3].

Each model was evaluated using **5-fold cross-validation** and tested on **80/20** and **50/50** train-test splits to ensure generalizability and stability across different sampling scenarios [4].

E. Model Evaluation

Each model's performance was assessed using a set of standard classification metrics to evaluate both predictive accuracy and robustness. Accuracy served as the primary metric, providing an overall measure of how well the models performed in classifying maternal risk levels [1][2]. To gain deeper insight into classification behavior across the three target categories—High, Mid, and Low—a confusion matrix was generated, highlighting the distribution of correct and incorrect predictions. Additionally, ROC curves using a one-vs-all strategy were plotted to evaluate the models' ability to distinguish between each risk level [4]. To ensure the reliability and generalizability of the results, a 5-fold cross-validation approach was employed. Furthermore, the models were tested on two data split configurations (80/20 and 50/50) to compare performance across varying proportions of training and testing data [4][5].

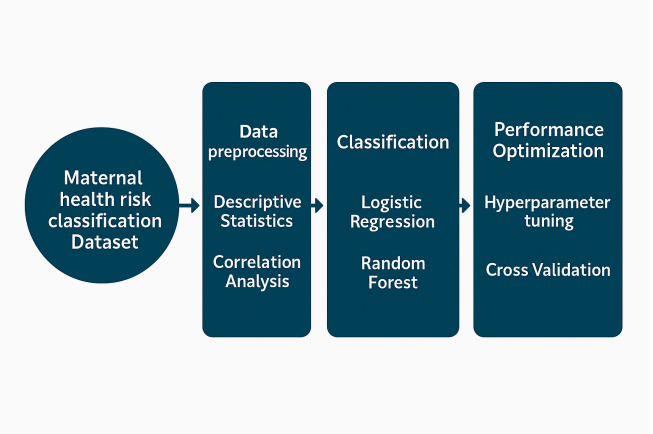


Figure 4.1 - Project Flow Chart

V. RESULTS AND ANALYSIS

In our project we trained two Machine Learning models Random Forest and Logistic Regression models using two distinct data-splitting methods: 80/20 (80% training, 20% testing) and 50/50 (equal training and testing). Evaluation metrics include accuracy, confusion matrices, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and ROC-AUC analysis.

Model	80/20 Split	50/50 Split
Random Forest	84.65%	80.27%
Logistic Regression	62.87%	63.7%

Random Forest model outperformed Logistic Regression in both the data splits indicating the data is non-linear [1], [4]. For the 80/20 data split Random Forest achieved 84.65% and for the 50/50% data split it achieved 80.27%. Whereas the Logistic Regression achieved 62.87% and for the 80/20% and 63.7% and for the 50/50%.

By observing the Confusion Matrix, Random Forest showed strong classification, especially for the "low risk" category and it is a bit confused while predicting the "low risk" category [5].

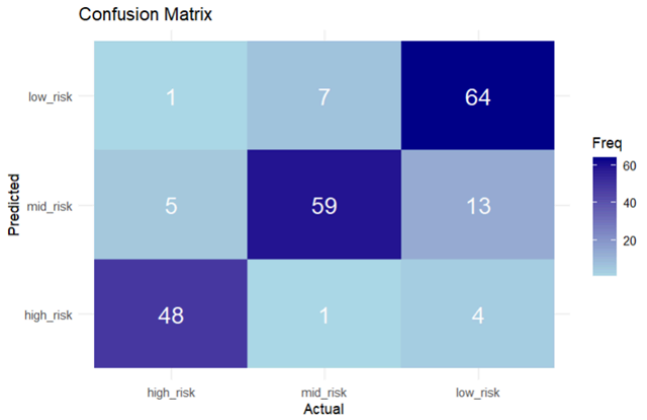


Figure 5.1.Random Forest Confusion Matrix for 80/20 Split

	Sensitivity	Specificity	PPV	NPV
Class: high_risk	0.8888889	0.9662162	0.9056604	0.9597315
Class: mid_risk	0.8805970	0.8666667	0.7662338	0.9360000
Class: low_risk	0.7901235	0.9338843	0.8888889	0.8692308

Evaluation Metrics of Random Forest Model for 80/20 Split

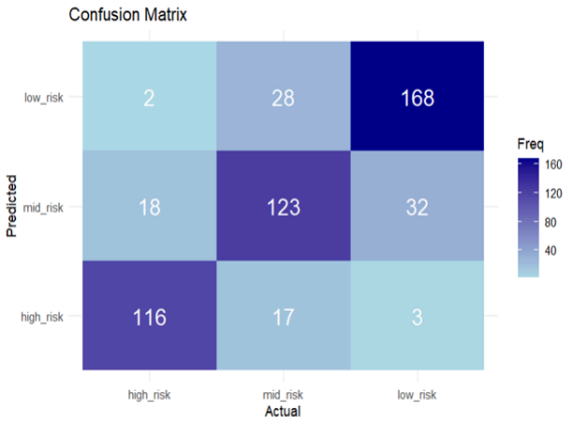


Figure 5.2. *Random Forest Confusion Matrix for 50/50 Split*

	Sensitivity	Specificity	PPV	NPV
Class: high_risk	0.8529412	0.9460916	0.8529412	0.9460916
Class: mid_risk	0.7321429	0.8525074	0.7109827	0.8652695
Class: low_risk	0.8275862	0.9013158	0.8484848	0.8867314

Metrics of Random Forest Model for 50/50 Split

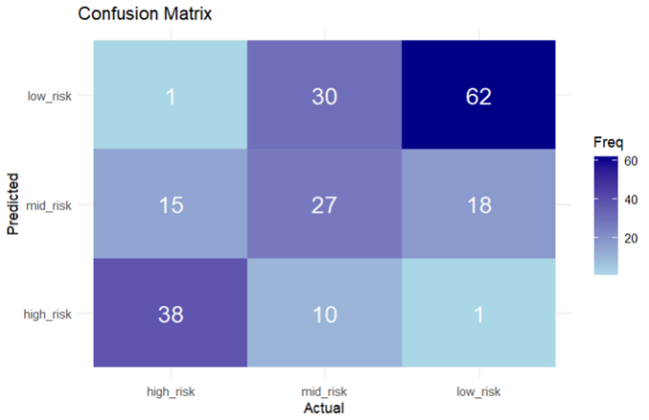


Figure 5.3. *Logistic Regression Confusion Matrix for 80/20 Split*

	Sensitivity	Specificity	PPV	NPV
Class: high_risk	0.7037037	0.9256757	0.7755102	0.8954248
Class: mid_risk	0.4029851	0.7555556	0.4500000	0.7183099
Class: low_risk	0.7654321	0.7438017	0.6666667	0.8256881

Metrics of Logistic Regression Model for 80/20 Split

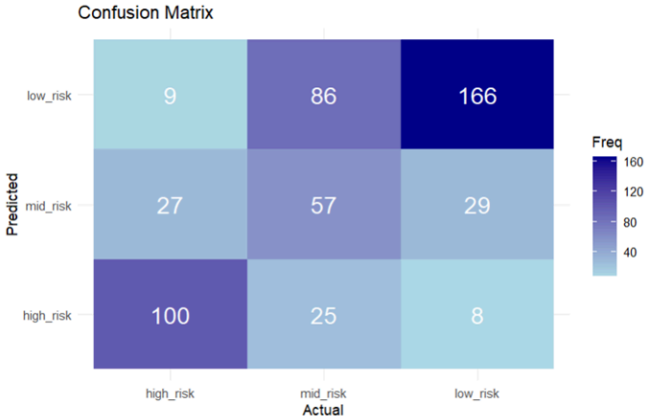


Figure 5.4. *Logistic Regression Confusion Matrix for 50/50 Split*

	Sensitivity	Specificity	PPV	NPV
Class: high_risk	0.7352941	0.9110512	0.7518797	0.9037433
Class: mid_risk	0.3392857	0.8348083	0.5044248	0.7182741
Class: low_risk	0.8177340	0.6875000	0.6360153	0.8495935

Metrics of Logistic Regression Model for 50/50 Split

ROC-AUC curves indicate that the Random Forest model predicted maternal risk levels accurately compared to the Logistic Regression. The ROC-AUC curve depicts how effectively the model differentiates between different risk categories. With an AUC of 0.979, the curves indicate excellent differentiations across all the risk levels, especially it indicates strong classification ability for "high risk" and "mid risk" categories. The curve for the "low risk" category shows slightly lower sensitivity, indicating a minor challenge in differentiating this class from others.

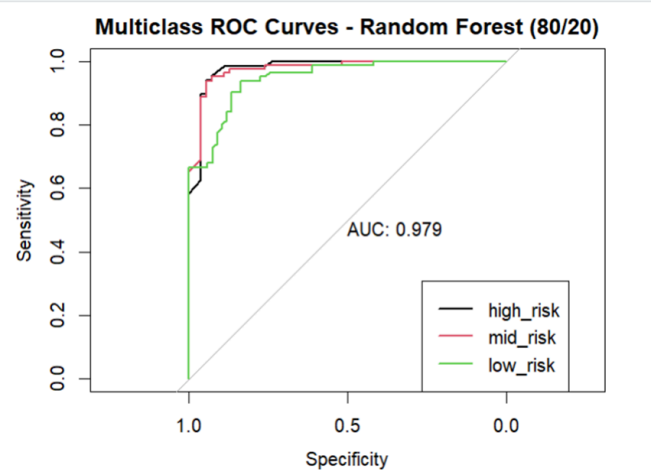


Figure 5.5 ROC Curve - Random Forest (80/20 Split)

Figure 5.5 shows the ROC curve for the Random Forest model with an AUC of 0.979, indicating excellent overall performance. The model effectively distinguishes "high" and "mid" risk categories, with slightly lower sensitivity for "low risk."

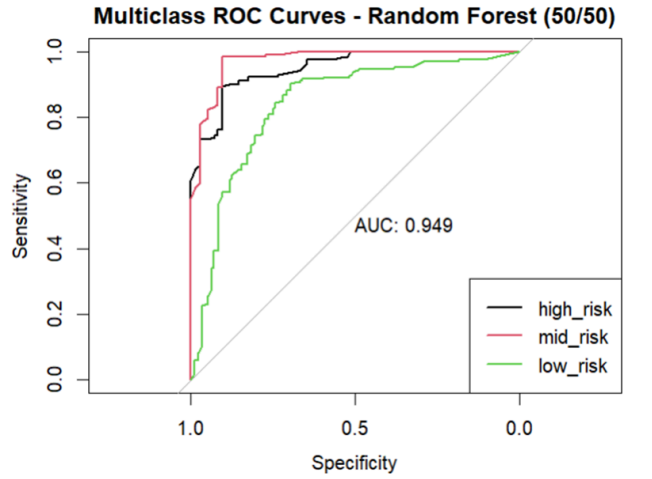


Figure 5.6 ROC Curve - Random Forest (50/50 Split)

Figure 5.6 presents the ROC curve for the Random Forest model with a 50/50 data split, achieving an AUC of 0.949. The model maintains strong classification performance, with slightly reduced sensitivity for the "low risk" class.

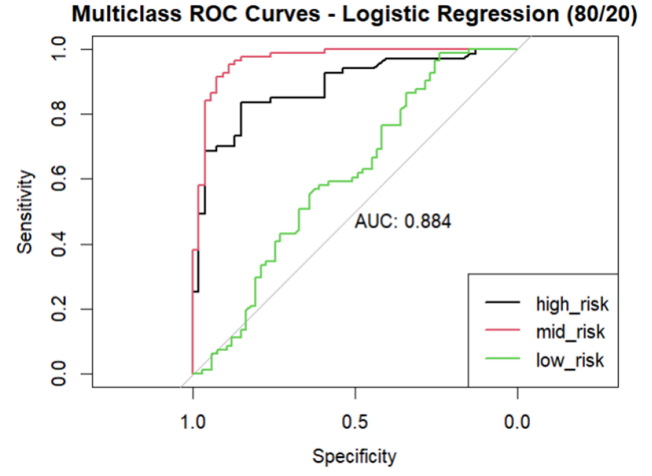


Figure 5.7 ROC Curve - Logistic Regression (80/20 Split)

Figure 5.7 displays the ROC curve for Logistic Regression using the 80/20 split, with an AUC of 0.884. While the model performs reasonably for "mid" and "high risk," it shows lower sensitivity for the "low risk" category.

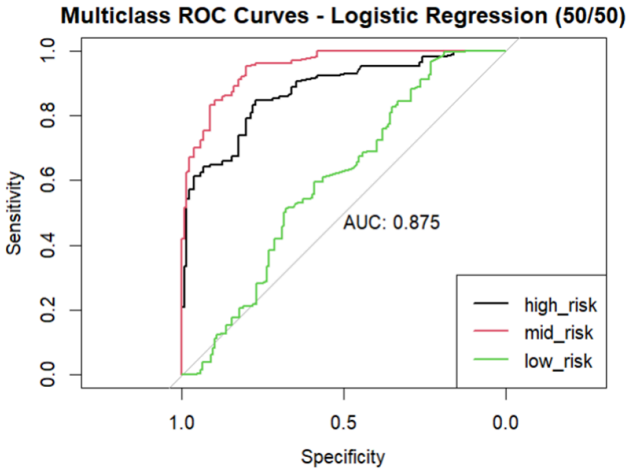


Figure 5.8 ROC Curve - Logistic Regression (50/50 Split)

Figure 5.8 illustrates the multiclass ROC curve for the Logistic Regression model evaluated on a 50/50 train-test split. The model achieved a strong overall performance with an AUC of 0.875, indicating high discriminative ability across the high, mid, and low risk classes.

The results clearly mention that the Random Forest model, with hyperparameter optimization and cross-validation, provides an accurate method for predicting maternal health risk levels during pregnancy period compared to Logistic Regression.

VI. DISCUSSION

Our results' analysis verifies that data mining techniques, particularly supervised machine learning models, can be effectively employed to ascertain the level of risk in instances of maternal health [1], [4]. Among all the models evaluated, Random Forest offered the most predictive accuracy, which bears witness to its potential for learning from patterns under health-related variables.

Our findings suggest that blood pressure (systolic and diastolic), age, and heart rate contribute significantly to the evaluation of the maternal risk category [6]. These results are also consistent with prevailing medical knowledge, where hypertensive disorders of pregnancy are considered to be leading risk factors during pregnancy.

This study lends support to the potential use of such predictive models in a clinical environment or as part of an integrated mobile healthcare application, particularly where experts are not easily available. With early warnings, these models can assist timely medical decisions and potentially prevent complications.

However, some limitations were observed. The dataset used did not include behavioral, environmental, or socioeconomic factors, which may influence the accuracy of risk prediction [2],[3]. The sample size was also moderate, and further testing on larger and more complex datasets is needed to determine the model's robustness.

Compared to other research available, our approach is a comparative study of various models and their effectiveness, with the added insight into how machine learning can be used in maternal healthcare [5]. For future work, more comprehensive health information and testing on actual datasets from hospitals or healthcare programs would enhance the performance and usability of the model.

VII. CONCLUSION

Maternal Health Risk Prediction is the most important research domain, which can greatly improve the health of the mothers. This study developed and evaluated Random forest and Logistic Regression Machine learning models to predict the Maternal Health risk levels after delivery, using the UCI Maternal Health Risk dataset. Random Forest has achieved a remarkable accuracy of 84.65% and 80.27% for 80/20 and 50/50 splits respectively. Logistic Regression with accuracies of 62.87% and 63.7% for 80/20 and 50/50 splits respectively.

These findings highlight that methods like Random Forest for predicting Maternal Health risk identify High-risk cases critical to maternal safety such as postpartum hemorrhage or severe maternal morbidity. However the lack of explicit postpartum specifications in the dataset limits direct applicability [6].

Future research should prioritize its focus on datasets with explicit postpartum outcomes, and which has feature importance analyses for clinical interpretability and explore more methods like gradient boosting or oversampling to enhance performance. Additionally, validating models on diverse populations is also essential for generalizability [7].

VIII. REFERENCES

[1] D. L. Gaur, P. Pandey, and R. Singh, "Predictive analysis of maternal health using supervised machine learning algorithms," International Journal of Scientific & Engineering Research, vol. 11, no. 6, pp. 678–685, 2020.

[2] M. J. Ahmed and A. K. Khan, "Data mining techniques for predicting maternal health risks," Journal of Biomedical Engineering and Medical Imaging, vol. 7, no. 4, pp. 45–52, 2020.

[3] A. Fatima and M. H. Yousaf, "An intelligent healthcare system for maternal risk prediction using data mining," Indonesian Journal of Electrical Engineering and Computer Science, vol. 17, no. 2, pp. 891–899, 2020.

[4] S. U. Rehman, F. Javed, and H. A. Shah, "Comparative study of machine learning models for risk prediction in maternal care," Asian Journal of Research in Computer Science, vol. 9, no. 3, pp. 23–31, 2021.

[5] R. Sultana and A. Begum, "Application of decision tree and SVM for maternal health data classification," Bangladesh Journal of Medical Informatics, vol. 13, no. 1, pp. 12–19, 2020.

[6] World Health Organization, "Trends in Maternal Mortality: 2000 to 2020," WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division, 2023.

[7] J. W. Lee, J. Y. Park, and Y. T. Park, "Predicting maternal risk factors in pregnancy using machine learning," Healthcare Informatics Research, vol. 24, no. 2, pp. 123–129, Apr. 2018

[8] Al Hinai, M. R., et al. (2024). Predicting maternal risk level using machine learning models. BMC Pregnancy and Childbirth. https://doi.org/10.1186/s12884-024-07030-9

[9] Oluwarotimi, A. A., et al. (2023). Deep hybrid model for maternal health risk classification in pregnancy: Synergy of ANN and Random Forest. Frontiers in Artificial Intelligence. https://doi.org/10.3389/frai.2023.1213436

[10] Pratama, R. W., et al. (2023). Prediction of maternal health risk levels based on multiple machine learning algorithms. ACE Conference Proceedings for the maternal risk prediction on ML. https://www.ewadirect.com/proceedings/ace/article/view/13675

[11] Patel, N., et al. (2023). Interpretable predictive models to understand risk factors for maternal and fetal outcomes. arXiv preprint. https://arxiv.org/abs/2310.10203

[12] Park, Y., et al. (2022). Comparison of machine learning and logistic regression as predictive models for adverse maternal and neonatal outcomes of preeclampsia. PLOS ONE. https://pubmed.ncbi.nlm.nih.gov/36312231

IX. PROJECT CONTRIBUTION

Contribution : Each member contributed equally

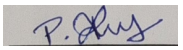
Meeting format : G-meet

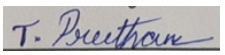
Group meet time and duration :

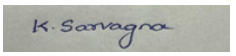
04/28/2025 – (2:00 PM – 6:00 PM)
&
04/29/2025 – (2:00 PM – 5:00 PM)

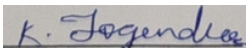
Average time in communication and discussion regarding assigned group work : 30 mins

Participants :

Roshan Pampari 

Preetham Thatikonda 

Sarvagna Kotte 

Yogendra Kodali 

Pavan Kumar Etta