Predicting Maternal Health Risk Using Machine Learning Algorithms

Boluwatife Olayiwola Elijah

Faculty of Engineering,

Environment and Computing Coventry University Coventry, England

Student ID: 12379264

olayiwolab@coventry.ac.uk

**ABSTRACT**

*As pregnancy and childbirth entail a considerable risk of difficulties that could have catastrophic repercussions for both the mother and the child, maternal health is a crucial topic of concern in healthcare. In this study, we use a sizable dataset of maternal health risk to apply three machine learning models—Decision Tree, XG Boost, and Gradient Boost. The dataset includes a variety of variables, including data on demographics, medical history, and pregnancies. We seek to determine the most accurate model for predicting maternal health risks and investigate the variables that are most closely related to those risks. Our findings demonstrate that the Decision Tree algorithm outperforms the other two models, with an F1 score of 78% and accuracy, precision, recall, and recall of 83%. Furthermore, our study shows that significant determinants of maternal health risks include age, medical history, and specific pregnancy-related issues such gestational diabetes. The outcomes of this study's findings may help clinical decision-making and enhance mothers’ health.*

**KEYWORDS**

*Maternal Health, Maternal Health Dataset, Risk Factor, Internet of Things, Machine Learning, Gradient Boosting Classifier, Extreme Gradient Boosting, Smote over sampling, Decision Tree.*

**INTRODUCTION**

As pregnancy and childbirth entail a considerable risk of difficulties that could have catastrophic repercussions for both the mother and the child, maternal health is a crucial topic of concern in healthcare it is crucial to identify maternal health hazards early in order to deliver timely interventions and improve outcomes. Machine learning algorithms can analyse patient data and identify risk variables to assist anticipate maternal health hazards. In this study, we predict maternal health risks in a large dataset using three well-known machine learning models: Decision Tree, XG Boost, and Gradient Boost. The dataset includes a variety of variables, including data on demographics, medical history, and pregnancies. By using these algorithms on the dataset, we hope to determine the most accurate model for predicting maternal health risks and investigate the variables that are most closely related to those risks. The outcomes of this study's findings may help clinical decision-making and enhance mother health.

**DATASET**

The dataset used in this study was obtained from the UCI Machine Learning Repository, which is a popular source of open-access datasets for research in machine learning. The maternal health risk dataset contains 1,014 instances, each representing a pregnant woman. The dataset contains 7 attributes that provide information about the demographic, medical, and pregnancy-related characteristics of the participants. The target variable is a binary indicator of maternal health risk, where 3 represents pregnancy a high-risk, 2 represents Mid risk pregnancy and 1 represents a low-risk pregnancy. The dataset has been pre-processed to remove missing values and perform feature scaling.

The dataset was originally collected to investigate the performance of different machine learning algorithms in predicting maternal health risks, with the aim of identifying important risk factors that could inform clinical decision-making. The dataset has been widely used in other studies as well, making it a valuable resource for researchers interested in maternal health risk prediction. By using an open-access dataset, this study is able to contribute to the larger machine learning community by providing insights into the performance of different algorithms in a real-world application. Overall, the use of an open-access dataset allows for greater transparency and reproducibility in research, which is essential for advancing the field of machine learning.

**Data Set Information**

Age, Systolic Blood Pressure as SystolicBP, Diastolic BP as DiastolicBP, Blood Sugar as BS, Body Temperature as BodyTemp, HeartRate and RiskLevel. All these are the responsible and significant risk factors for maternal mortality, one of the main concerns of SDG of UN.

**Attribute Information**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Set Characteristics | N/A | Number of Instances | 1014 | Area | Life |
| Attribute Characteristics | N/A | Number of Attributes | 7 | Date Donated | 2020 – 12 - 31 |
| Associated task | Classification | Missing Values | N/A | Number of web hits | 26133 |

*Age: Any ages in years when a women during pregnant.  
SystolicBP: Upper value of Blood Pressure in mmHg, another significant attribute during pregnancy.  
DiastolicBP: Lower value of Blood Pressure in mmHg, another significant attribute during pregnancy.  
BS: Blood glucose levels is in terms of a molar concentration, mmol/L.  
HeartRate: A normal resting heart rate in beats per minute.  
Risk Level: Predicted Risk Intensity Level during pregnancy considering the previous attribute.*

**METHODOLOGY**

The project's main goal is to identify, assess, and predict maternal health concerns utilising a variety of machine-learning feature selection and classification techniques. In order to determine which, set of features and classifiers provides the best accurate result for forecasting maternal health risks, also present and contrast findings using feature selection and feature-free applications to various classifiers. We apply the classifiers to the data set we've chosen and evaluate their performance using accuracy scores. The Dataset was Imported, cleaned, and engineered data from the UCI repository to predict which patients have a higher risk of experiencing complications from pregnancy using different classification models as shown below:

**Data exploration**

The dataset was critically analysed to understand its properties, relationships and patterns. Getting insights into the data that can guide decision-making and inspire additional investigation is the aim of data exploration. It helped to detect any problems with the data, a better knowledge of the data and pinpointed the data's key characteristics, which guided the selection of features for use in predictive model construction.

After the data exploration it was observed that there was a class imbalance in the data set, “the visualization below shows the imbalanced in the predictor used”. To solve this problem the following method was used:

Fig 1. Class imbalance

**SMOTE OVERSAMPLING:**

The dataset was split into two (x, y) and Smoot oversampling was used to balance the class distribution within a dataset. By underrepresenting the minority class in the dataset, this strategy attempts to correct the problem of class imbalance that results in biased models. From the chart it is shown that the number of samples in 2 (mid risk) is much smaller

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Features | Age | Systolic Bp | Diastolic Bp | BS | Body Temp | Heartrate |
| Mean of features before scaling | 29.871795 | 113.198225 | 76.460552 | 8.725986 | 98.665089 | 74.301775 |
| Standard deviation of features before scaling | 13.474386 | 18.403913 | 13.885796 | 3.293532 | 1.371384 | 8.088702 |
| Mean of features after scaling | -4.90512737e-17 | -7.70805729e-17 | 4.69490762e-16 | -2.13723407e-16 | 2.15124872e-15 | 1.36642834e-16 |
| Standard deviation of features after scaling | 1 | 1 | 1 | 1 | 1 | 1 |

than the number of samples in the other classes 1 (low risk), and 3 (high risk), leading to a biased model. By using SMOTE, the number of samples in the minority class can be increased, making the dataset more balanced, and leading to a more accurate model. Below shows after the class balancing:

Icon

Description automatically generated

**Dataset Scaling**

Dataset scaling is a pre-processing method used in machine learning to transform the scale and distribution of the features in a dataset. This is crucial as many machine learning algorithms are sensitive to the size and distribution of the features and may underperform if these factors are ignored.

There are several methods for scaling a dataset, but standard scaling was used on the dataset.

***Standard scaling:***

*Standard scaling transformed the dataset's features to give them a mean of 0 and a standard deviation of 1. This aids in removing the influence of various feature sizes and distributions, which can result in inaccurate predictions.*

*A dataset must be scaled for a variety of reasons. First of*

*all, it can aid in preventing specific features from having an unbalanced influence on the model, which could result in skewed predictions. Second, it can aid in making sure the features are in a format appropriate for usage with numerous machine learning algorithms, which can result in higher performance and more precise predictions. Finally, since the data is simpler for the algorithm to deal with, it can aid in accelerating the training process for some algorithms.*

**Machine Learning Algorithms Used**

Machine learning techniques known as classifiers are employed to forecast a target variable based on a collection of input information. Classifiers are employed in the context of maternal health risk prediction to determine the risk level of expectant mothers based on data including demographics, medical history, and risk factors related to pregnancy.

In this study, the following classifiers were employed:

**Decision Tree:**

The decision tree classifier constructs a tree-like model, where each leaf node stands in for the anticipated class label and each inside node represents a feature or characteristic. The method begins at the root node and examines the values of the features for a given case before moving up the branch of the tree that corresponds to the feature that is most important to the case until it reaches a leaf node. The ultimate prediction in that situation is the class label given to that leaf node.

By analysing the relationship between these features, the algorithm can learn to identify patterns and make predictions about the risk of a health complication in future pregnancies.

Overall, due to its simplicity of interpretation and comprehension, the decision tree classifier was a beneficial tool for identifying threats to maternal health getting an accuracy score of 80% above all other classifiers used on the dataset.

Below shows the confusion matrix for decision tree classifier and a visualization for the decision tree.

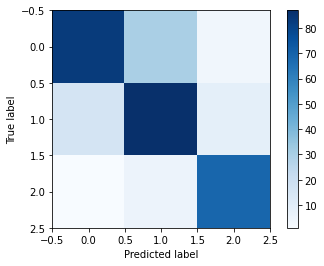


Fig 2 confusion matrix for decision tree classifier

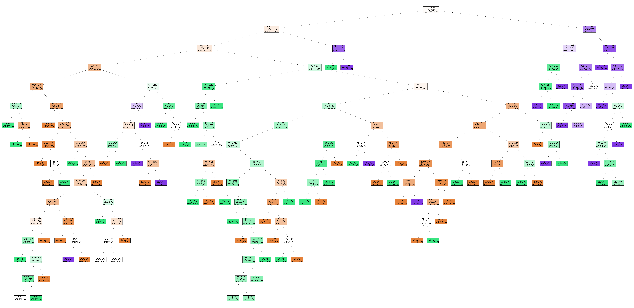


Fig 3 Decision tree.

**XG-Boost Classifier:**

XG-Boost is a popular and effective machine learning algorithm for classification problems and was a useful tool for predicting maternal health risks in the dataset.

One benefit of XG-Boost is its capacity to properly handle missing data as well as continuous and categorical data which makes it an important classifier for the Dataset being worked on, getting an accuracy of approximately 72%.

Below shows the confusion matrix and a visualization for XG-Boost classifier (The function displays the relative relevance of each model element, making it possible to decide which features are most crucial for making predictions.).

Calendar

Description automatically generated

Fig 4

Square

Description automatically generated

Fig 5 confusion matrix for XG-Boost classifier

**Gradient Boost Classifier:**

With the aid of Gradient boosting, it was possible to determine whether a pregnant woman is in danger of developing a certain health issue, such as pre-eclampsia, gestational diabetes, or preterm labour.

Gradient Boosting has the advantage of being able to handle both continuous and categorical data as well as successfully manage missing data which makes it an important classifier for the Dataset being worked on, getting an accuracy of approximately 75%.

**Model evaluation metrics:**

* Model Accuracy
* Model Precision
* Recall
* F1 Measure

**MODEL TABLE RESULTS AND EXPLANATIONS:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithms | Accuracy | Precision | Recall | F1 |
| Gradient Boost | 0.75 | 0.75 | 0.73 | 0.74 |
| XG-Boost | 0.72 | 0.70 | 0.76 | 0.73 |
| Decision Tree | 0.80 | 0.83 | 0.74 | 0.78 |

*Decision Tree:*

*The model's accuracy was 80%, meaning that 80% of the time it accurately predicted the target class. It is the highest of the three types.*

*XG-Boost:*

*When compared to the Decision Tree, the model's accuracy is lower at 72%. This indicates that the model may not perform well in some situations, as it properly identified the target class in 72% of the cases.*

*Gradient Boost:*

*The model is more accurate than XG-Boost but less accurate than Decision Tree, with a 75% accuracy rate. This indicates that the model performed better than XG-Boost but not always as well as Decision Tree, correctly predicting the target class in 75% of the situations.*

**Conclusion**

The Decision Tree approach outperforms the other two, according to the analysis of the three algorithms' performance on the dataset for maternal health risk. However, this judgement is supported by the measurements for accuracy, precision, recall, and F1 score. To elaborate further, the following ideas can be taken into account:

***The balance between precision and recall:***

Of the three algorithms, Decision Tree gets the greatest F1 score, indicating a strong balance of precision and recall. Recall represents the percentage of true positive examples that the model properly identifies, whereas precision is the percentage of positive predictions that are genuinely true positive. A high F1 score indicates that the model has a good balance between recall and precision, i.e., it can recognise the majority of positive examples while still making accurate positive predictions.

***The trade-off between false positive and false negative:***

In some situations, precision is more crucial than memory, and vice versa. A false negative, such as identifying a high-risk mother as low-risk, may have catastrophic repercussions, whereas a false positive, such as identifying a low-risk mother as high-risk, would necessitate unnecessary measures. A model with a higher recall would be desired in this scenario.

***Computing complexity:***

While Gradient Boost is a more sophisticated algorithm that can take longer to train, Decision Tree and XG-Boost are comparatively basic algorithms that are computationally efficient. The size of the dataset, the available computing power, and the intended training duration all influence the algorithm that is used.

***Interpretability:***

Decision Tree provides a tree-like structure to express the relationship between the features and the target. It is an interpretable method. As a result, it is simpler to comprehend how the model generated its predictions. It is more difficult to comprehend how the model arrived at its predictions because XG-Boost and Gradient Boost are less interpretable. When using results in a clinical or regulatory environment, interpretability can be a crucial element to take into account.

*In conclusion, the Decision Tree algorithm performed the best in terms of accuracy, precision, recall, and F1 score; however, before choosing the best algorithm for the task, it is important to take other factors into account, such as computational complexity, interpretability, and trade-offs between false positive and false negative.*

**Reference**

1. Ahmed M., Kashem M.A., Rahman M., Khatun S. (2020) Review and Analysis of Risk Factor of Maternal Health in Remote Area Using the Internet of Things (IoT). In: Kasruddin Nasir A. et al. (eds) InECCE2019. Lecture Notes in Electrical Engineering, vol 632. Springer, Singapore. [Web Link]

2. IoT based Risk Level Prediction Model for Maternal Health Care in the Context of Bangladesh, STI-2020, [under publication in IEEE]

3. <https://github.com/Gourang1804/Maternal-Health-Risk-Prediction>

4. <https://github.com/daanishgoyal/Maternal-Health-Risk-Prediction/blob/main/README.md>

**Codes And Dataset Links**

Code

<https://github.com/BOLUWATIFE-OLAYIWOLA/Maternal-Health-Risk-Dataset-Codes>

Dataset

<https://archive.ics.uci.edu/ml/datasets/Maternal%20Health%20Risk%20Data%20Set#>