Prediction for Patient Maternal Health Risk Level

Manikanta Katuri

Seanmark Paz

Chris A

Nabeel Khan

Department of Applied Artificial Intelligence

AAI-501-01: Introduction to Artificial Intelligence

Andrew Vaughan

12-Aug-2024

# **Abstract**

Prediction for Maternal Health risk for patients or the maternal outpatients provides illustration of various statistical and machine learning methods to predict the risk level for the patients. The primary concern addressed in this paper about improving maternal health, reducing maternal and child mortality. This paper demonstrates effective monitoring of pregnant women mostly in rural areas of Bangladesh, which indicates the pregnant women and their family about the health conditions. Many health factors can influence maternity and some among them like age of the patient, blood pressure, blood sugar levels, body temperature, heart rate are considered in this study for analysis and prediction. The dataset collected from different hospitals and community clinics of many rural areas of Bangladesh have these features and the pregnancy risk factors are predicted. The dataset obtained from the UCI machine learning library contains 1013 enrollments of the maternal patients with these features: Ahmed,Marzia. (2023). Maternal Health Risk. UCI Machine Learning Repository. <https://doi.org/10.24432/C5DP5D> . This dataset is fit into machine learning model by splitting into training and test datasets. Different machine learning classifiers are considered for evaluation and the GridSearchCV decision tree classifier has provided the maximum accuracy score for the dataset yielding a mean success prediction rate of 83% and some of the predictions are more than 90% accurate.

# **Introduction**

Infant mortality has been a major concern in many underdeveloped or developing countries even to this era. Many pregnant women lose their lives with severe health conditions, and this is due to lack of proper medical monitoring systems and information about maternal health care. Pregnant women must have regular health monitoring and care to assist with healthy living of the women and child. Some of the external factors those can influence health of pregnant women are: mental health conditions influencing factors including work environment, nutrition, air quality, water quality, living conditions, excessive exposure to sun’s heat in the Asian or middle-eastern countries causing dehydration and other blood pressure related conditions. These conditions are to be monitored for optimal living environment and healthy mental and physical state of both the pregnant women and fetus. Unfortunately, some of the rural regions of developing countries do not have facilities to improve the knowledge and providing timely health related recommendations to these pregnant women therefore causing abnormal deliveries impacting the child and pregnant women mortality.

Smaller Internet of Things (IoT) devices can be of greater help in this scenario those are highly capable of collecting the health features and transmitting to other local servers and computers, and the data from these features can be analyzed to know the current health conditions and predict the maternal risk status. A prototype of the data collection, analysis process and decision-making algorithms predicting the health risk can be represented as in the Figure 1.

**Figure 1**

*Prototype of the IoT and ML algorithms interaction to predict health*

A diagram of a data source

Description automatically generated

This research is to solve the problem of infant mortality and death risk of the pregnant women in the under-developed and developing countries. The first problem is the collection of data. This can be achieved through wearable IoT devices or devices at the medical facilities connected over internet and those can transmit data into larger data collecting systems like cloud or local servers. The second problem is the accurate analysis and prediction of the health risk. Various Machine learning classification algorithms like SVM (Support Vector Machines), Decision tree, GSV (Grid Search Classification Vector), Randomized Search CV are considered and implemented for data analysis and building a model that has the highest accuracy. Decision trees are the most advanced and helpful algorithms for accuracy and prediction rather than the regression models. The third task is to provide health professionals with a diagram which increases the probability of accurate patient risk classification that balances accuracy and simplicity.

# **Exploratory Data Analysis**

The dataset obtained at the UCI Machine Learning library consists of the following features: Age of the patient, Systolic blood pressure, Diastolic blood pressure, Blood sugar levels, body temperature, heart rate. The risk level of the patient health is classified into low, mid and high considering the values of the features. A sample of the data in this dataset is shown in Figure 2.

**Figure 2**

*Sample of the data in the initial dataset*

A table with numbers and text

Description automatically generated

Several cross tabulations were performed to determine what variables could be used for classifying the health risk of the patient.

**Figure 3**

*Statistical data showing risk levels of patients for different body temperatures*

A screenshot of a graph

Description automatically generated

**Figure 4**

*Statistical data showing risk levels of patients for different heart rates*

A table of numbers with numbers on it

Description automatically generated

**Figure 5**

*Statistical data showing risk levels of patients for different systolic and diastolic blood pressures*

A table of numbers with numbers

Description automatically generated A table with numbers and letters

Description automatically generated

**Figure 6**

*Statistical data showing risk levels of patients with different blood sugar levels*

A table with numbers on it

Description automatically generated

The statistical data shown above has no significance in indicating any contributing features for high risk of the patients. Hence, the data is further considered for input to various machine learning classifying algorithms to find the accuracy and obtaining a decision tree.

### **Data Cleaning/Preparation**

# The data is categorized as two equal halves of test and training data. Groups of data is fit into SVM, Decision tree, grid search optimized decision tree, and randomized search optimized decision tree through python code. The grid search and randomized search methods were utilized to optimize the parameters for the decision tree model. The parameter of max\_depth was tuned in a way that would balance both accuracy and simplicity of the model. On average, the most optimized decision tree classifier would have an accuracy of around 83%.

# **Model Analysis**

The proposed model algorithm implements the following strategy:

Inputs: Age, BodyTemp, HeartRate, SystolicBP, DiastolicBP, BS(Blood Sugar)

Output: low / mid / high risk level for a patient.

The pseudocode for these machine learning algorithms is as follows:

***BEGIN***

***Read the input file and loop through each patient record***

***LOOP***

***Choose features from the patient record***

***Calculate significance for each feature, calculate accuracy and error rate***

***Implement statistical data approach and calculate the best value for a combination***

***For other combination of features repeat the loop***

***END LOOP***

***COMPARE with trained data and test data***

***Predict the output risk level and display the values***

***END***

Before the focus on the decision tree classifier, a comparison was made with the SVM classifier. The results of that can be seen in Figure 7 below, and it shows a clear winner in the decision tree when it comes to accuracy.

**Figure 7**

*Prediction accuracy score comparison between SVM and Decision tree classifiers*

A graph of different colored shapes

Description automatically generated with medium confidence

After deciding on using the decision tree as the model of choice, multiple techniques were utilized to optimize its accuracy. One of these techniques was optimization of the max\_depth parameter. As shown in Figure 8, model accuracy increases with model depth up until around 15, and it plateaus around 80/90 percent for the training and testing data, respectively.

**Figure 8**

*Model accuracy at different max\_depth values*

A graph with blue and orange lines

Description automatically generated

Two methods called the grid search and randomized search were also used to find the best hyperparameter values for the decision tree classifier. After running through 100 different models for each method, the scores were tallied and can be seen in Figure 9. The grid search method produced better results between the two with an average accuracy of 83%.

**Figure 9**

*Prediction accuracy score comparison between grid search and randomized search methods*

A graph of different colored bars

Description automatically generated

In addition to the hyperparameters optimized with the grid search method, ccp\_alpha was also optimized by training the classifier with a range of different ccp\_alpha values. Figure 10 shows the relationship between the different ccp\_alpha values and accuracy.

**Figure 10**

*Model accuracy at different ccp\_alpha values*

A graph with blue and orange lines

Description automatically generated

After finding the optimal ccp\_alpha value, a feature ranking was also generated from the grid search method. To reduce model complexity, only the top three features were chosen to prune the tree. Figure 11 displays the decision tree diagram of the model with optimized hyperparameters and feature restriction. The tree has an accuracy score of about 80%. However, this model still seemed a bit too complex for use in a clinical setting. Therefore, the decision was made to limit the max depth of the tree to 4, which can be seen in Figure 12. While this did result in a reduction in accuracy from about 80% to 71%, the reduced complexity of the tree allows for much easier use.

**Figure 11**

*Decision tree diagram with optimized hyperparameters*

A computer network diagram of a tree

Description automatically generated with medium confidence

**Figure 12**

*Decision tree diagram with a reduced max\_depth*

A diagram of a company

Description automatically generated

# **Conclusion and Recommendations**

# **References**

1. M. Ahmed and M. A. Kashem, "IoT Based Risk Level Prediction Model For Maternal Health Care In The Context Of Bangladesh," 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2020, pp. 1-6, doi: [10.1109/STI50764.2020.9350320](http://dx.doi.org/10.1109/STI50764.2020.9350320).
2. Mondal, S., Nag, A., Barman, A.K., & Karmakar, M. (2023). Machine Learning-based maternal health risk prediction model for IoMT framework. International Journal of Experimental Research and Review. [[PDF] Machine Learning-based maternal health risk prediction model for IoMT framework | Semantic Scholar](https://www.semanticscholar.org/paper/Machine-Learning-based-maternal-health-risk-model-Mondal-Nag/8e3e150c98921e412129476ef07630338d6d5287)
3. D. Bajaj, R. Kumari and P. Bansal, "Risk Level Prediction for Maternal Health Using Machine Learning Algorithms," 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), Greater Noida, India, 2023, pp. 405-409, doi: [10.1109/ICCSAI59793.2023.10421156](https://ieeexplore.ieee.org/document/10421156).

# **Appendix**