Maternal Health Risk Prediction

A Comparative Study of Machine Learning Model Performance

Mengen Liu, Roy Oh, Kim Tan Palanca, & Nicolas Zhu

Summary

This project aims to use machine learning to predict maternal health risk based on key physiological factors by comparing the performances between two models. Maternal health are a major concern in healthcare. Hence, early risk assessment is crucial for the welfare of both the mother and child. By leveraging structured data from the Maternal Health Risk Dataset, we trained and evaluated machine learning models to classify maternal health risk levels into three different classes: low, mid, and high risk.

The dataset contains 1,014 records with features such as age, blood pressures, blood sugar levels, body temperature and heart rate. Our goal will be to determine whether machine learning algorithms could accurately predict risk levels based on these physiological markers. We will first conduct an exploratory data analysis before building any of the models. Once we have trained the models, their accuracies will be evaluated as measures of performance.

Introduction

Maternal health can be defined as "the health condition of women during preganancy, childbirth, and the postnatal period (WHO 2025). This is a critical area of healthcare, as complications during preganancy and childbirth can lead to severe consequences for both mothers and newborns. According to the (WHO 2024), around 800 women died each day in 2020 due to preventable causes related to maternal health, further emphasizing the need for risk assessment measures.

Historically, risk assessment have been carried out by medical professionals that relied heavily on clinical expertise and constant monitoring. However, traditional approaches to monitoring basic physiological indicators often lacked efficiency in identifying potential complications (Mu, Yan, and Zhu 2023). Since the boom of machine learning (ML), many members of the academe have explored the use of ML in maternal health risk prediction, offering data-driven approaches to enhance early detection and intervention to offload the burden on overworked medical professionals (Bajaj, Kumari, and Bansal 2023; Mu, Yan, and Zhu 2023; Ukrit et al. 2024).

The analysis will utilize the Maternal Health Risk dataset sourced from the UC Irvine Machine Learning Repository (Ahmed 2020). Consisting of 1014 observations, this dataset includes the following 7 features:

- `Age`: Age of the patient (in years).
- `SystolicBP`: Systolic Blood Pressure (mmHg).
- `DiastolicBP`: Diastolic Blood Pressure (mmHg).
- `BS (Blood Sugar Level)`: Blood sugar concentration (mmol/L).
- `BodyTemp`: Body temperature (°F).
- `HeartRate`: Heart rate (beats per minute).
- `RiskLevel`: The target variable, categorized into low risk, mid risk, and high risk.

Project Question

To contribute to this discourse, this research aims to conduct a comparative study on the performance of two ML techniques in predicting maternal health risk, assessing each model's reliability in identifying risk levels. The following research question guides this analysis:

- *Which machine learning model most accurately predicts the maternal health risk level (low, medium, or high risk) based on physiological indicators such as blood sugar levels, body temperature, and other relevant health factors?*

Methods

For this analysis, the data will first be loaded into the notebook then cleaned to handle any possible missing values and ensure its usability for the various models. Following the data cleaning stabe will be an exploratory data analysis (EDA) to gain a comprehensive view of the data. This step will include visualizing the summary statistics, distributions, and correlations between variables to determine any patterns in the data prior to the model development.

This study will implement 3 ML classification models:

1. Baseline (Majority Class)
2. Logistic Regression
3. Random Forest

Each model will be evaluated based on the appropriate classification metric to compare their relative performance in maternal health risk prediction.

Table 1: Check for Missing Values

```
# A tibble: 7 x 2
  feature        na
  <chr>       <dbl>
1 Age             0
2 SystolicBP      0
3 DiastolicBP     0
4 BS              0
5 BodyTemp        0
6 HeartRate       0
7 RiskLevel       0
```

Table 2: Distinct Risk Levels

```
# A tibble: 3 x 1
  RiskLevel
  <chr>
1 high risk
2 low risk
3 mid risk
```

Wrangling and Cleaning the Data

From Table 1, we find that there are no missing values in the dataset.

Furthermore, we find that the features `Age`, `SystolicBP`, `DiastolicBP`, `BS`, `BodyTemp`, and `HeartRate` are numeric variables, while `RiskLevel` is currently a character variable. Moreover, Table 2 shows that there are three categories under `RiskLevel`: high risk, mid risk, and low risk. Given the three distinct categories under the target feature, we will modify `RiskLevel` to a factor variable to appropriately reflect its categorical nature in further analysis.

Table 3 shows a snapshot of the cleaned data.

Table 3: Sample of the Cleaned Data

```
# A tibble: 6 x 7
    Age SystolicBP DiastolicBP    BS BodyTemp HeartRate RiskLevel
  <dbl>      <dbl>       <dbl> <dbl>    <dbl>     <dbl> <chr>
1    25        130          80 15          98        86 high risk
2    35        140          90 13          98        70 high risk
3    29         90          70  8         100        80 high risk
4    30        140          85  7          98        70 high risk
5    35        120          60  6.1        98        76 low risk
6    23        140          80  7.01       98        70 high risk
```

Table 4: Summary Statistics

```
# A tibble: 6 x 6
  Feature       min   max   mean median    sd
  <chr>       <dbl> <dbl>  <dbl>  <dbl> <dbl>
1 Age            10    70  29.9    26   13.5
2 SystolicBP     70   160 113.    120   18.4
3 DiastolicBP    49   100  76.5    80   13.9
4 BS              6    19   8.73    7.5  3.29
5 BodyTemp       98   103  98.7    98    1.37
6 HeartRate       7    90  74.3    76    8.09
```

Exploratory Data Analysis

Summary Statistics

Age Distributions

Since age is an important factor in maternal health, we visualize the age distribution by risk level. From the visualization, high risk individuals have a higher median age around 35 years old. Additionally, the interquartile range indicates that the high risk group has more variation in age. We observe some outliers in the low and mid risk groups. Based on the visualization, older aged individuals seem more associated with maternal health risks.

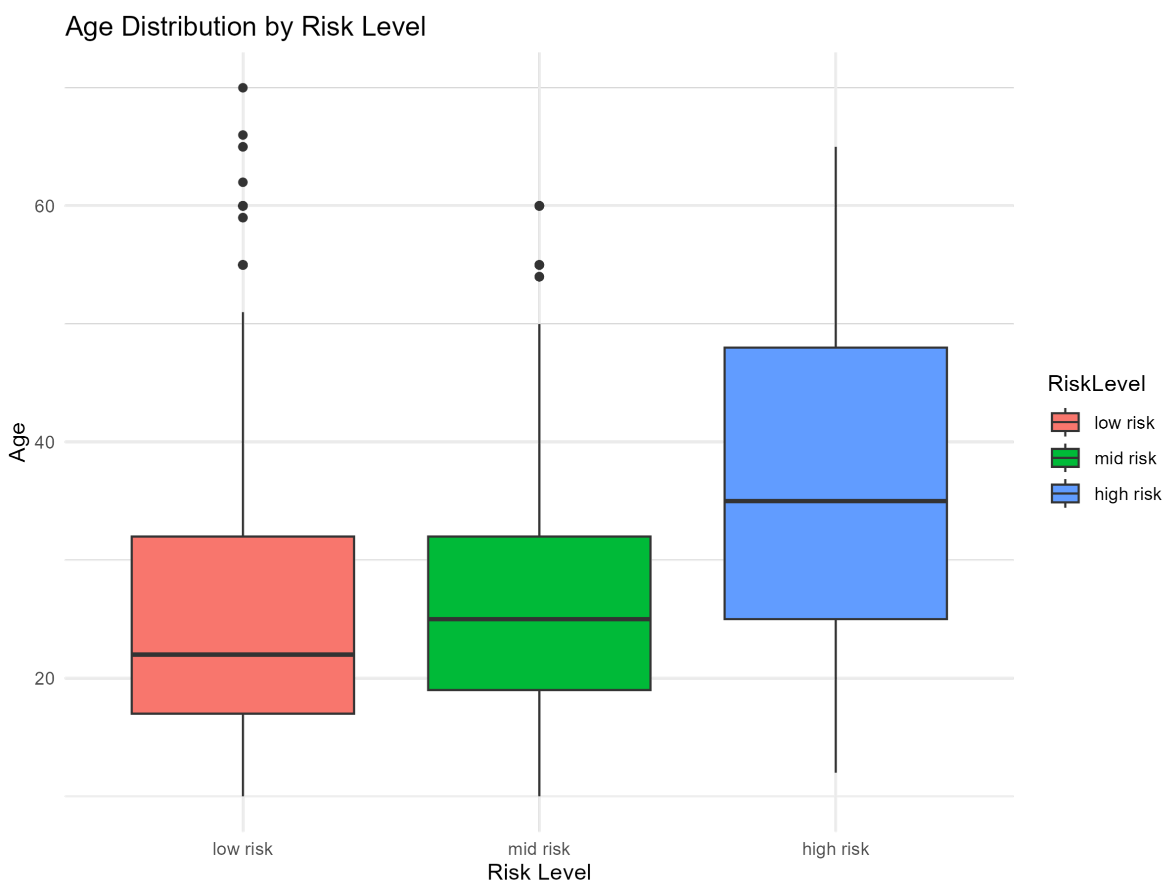
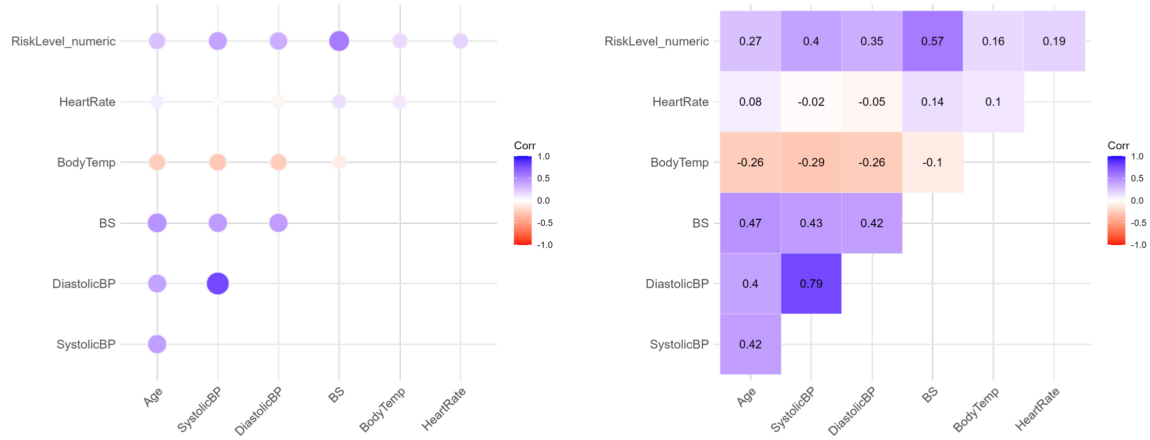


Figure 1: Age Distribution by Risk Level

Since the target classes seem relatively balanced, it would be appropriate to use **accuracy** as the main scoring metric. Accuracy is given by the number correct prediction out of all predictions made

Correlation Matrix



(a) Ellipse Correlation

(b) Correlation Values

Figure 2: Correlation Matrix

All of the variables have a positive correlation with RiskLevel, indicating that increases in these variables generally correspond to a higher maternal health risk. BS (Blood Sugar level) has the strongest correlation of 0.7870065, suggesting it is likely to be the most influential factor. We thought age would have a stronger correlation with RiskLevel, however, systolic blood pressure and diastolic blood pressure seems to have a stronger correlation with RiskLevel than age. Additionally, there are no signs of concern for multicollinearity issues in this dataset.

These findings may be a possible reason for the outliers observed above. Younger individuals with high blood pressures or sugar levels may be classified into higher risk levels. This indicates the importance of other factors.

Classification Model Building

Train/Test Splitting

The cleaned data will be split into training and testing sets, with 80% allocated for training and 20% for testing.

Baseline Model (Majority Class)

The baseline model has shown an accuracy of 0.4009901.

Table 5: Multinomial Logistic Regression Model

```
Call:
multinom(formula = RiskLevel ~ ., data = train_data)

Coefficients:
          (Intercept)          Age SystolicBP DiastolicBP        BS  BodyTemp
mid risk    -48.31965 -0.007544419 0.05929424 -0.03362508 0.3747446 0.3986736
high risk   -95.40388 -0.023851313 0.06420403  0.02026145 0.7873117 0.7628112
           HeartRate
mid risk  0.03010820
high risk 0.06461374

Residual Deviance: 1246.411
AIC: 1274.411
```

Multinomial Logistic Regression

The MLR model is as follows:

Model Testing

The multinomial logistic regression has only given us an slightly better accuracy than the baseline with a score of 0.6138614.

Lastly, we plot multinomial logistic regression graphs to visualize how predicted probabilities changes among different variable levels.

We observe from Figure 3 that as blood sugar level rises the probability of high risk increases, while mid and low risk decreases.

Random Forest

The random forest model is as follows:

Here are the parameters that have been passed through the randomForest object:

- `RiskLevel ~ .`: Predicts RiskLevel based on all other features
- `ntree = 500`: Uses 500 trees in the forest
- `importance` = TRUE: Computes feature importance

Table 6: Exponentiated coefficients to transform log-odds to odds ratio

```
# A tibble: 14 x 6
   y.level   term        estimate std.error   statistic   p.value
   <chr>     <chr>          <dbl>     <dbl>       <dbl>     <dbl>
 1 mid risk  (Intercept) 1.04e-21  0.000176 -274453.    0
 2 mid risk  Age         9.92e- 1  0.00774       -0.975 3.29e-  1
 3 mid risk  SystolicBP  1.06e+ 0  0.00813        7.30  2.98e- 13
 4 mid risk  DiastolicBP 9.67e- 1  0.0106        -3.18  1.48e-  3
 5 mid risk  BS          1.45e+ 0  0.0870         4.31  1.67e-  5
 6 mid risk  BodyTemp    1.49e+ 0  0.0130        30.6   2.72e-205
 7 mid risk  HeartRate   1.03e+ 0  0.0124         2.43  1.50e-  2
 8 high risk (Intercept) 3.69e-42  0.000147 -650390.    0
 9 high risk Age         9.76e- 1  0.0123        -1.95  5.16e-  2
10 high risk SystolicBP  1.07e+ 0  0.0121         5.31  1.12e-  7
11 high risk DiastolicBP 1.02e+ 0  0.0153         1.32  1.86e-  1
12 high risk BS          2.20e+ 0  0.0929         8.48  2.32e- 17
13 high risk BodyTemp    2.14e+ 0  0.0172        44.3   0
14 high risk HeartRate   1.07e+ 0  0.0167         3.87  1.10e-  4
```

Table 7: Multinomial Logistic Regression Predicted Probabilities

```
# A tibble: 202 x 5
      ID Predicted_Class `low risk` `mid risk` `high risk`
   <dbl> <chr>                <dbl>      <dbl>       <dbl>
 1     1 high risk           0.0131      0.204     0.783
 2     2 low risk            0.483       0.364     0.153
 3     3 mid risk            0.257       0.636     0.107
 4     4 low risk            0.509       0.402     0.0892
 5     5 low risk            0.652       0.249     0.0990
 6     6 low risk            0.893       0.103     0.00394
 7     7 mid risk            0.326       0.596     0.0778
 8     8 low risk            0.626       0.315     0.0595
 9     9 low risk            0.664       0.203     0.133
10    10 low risk            0.818       0.149     0.0333
# i 192 more rows
```

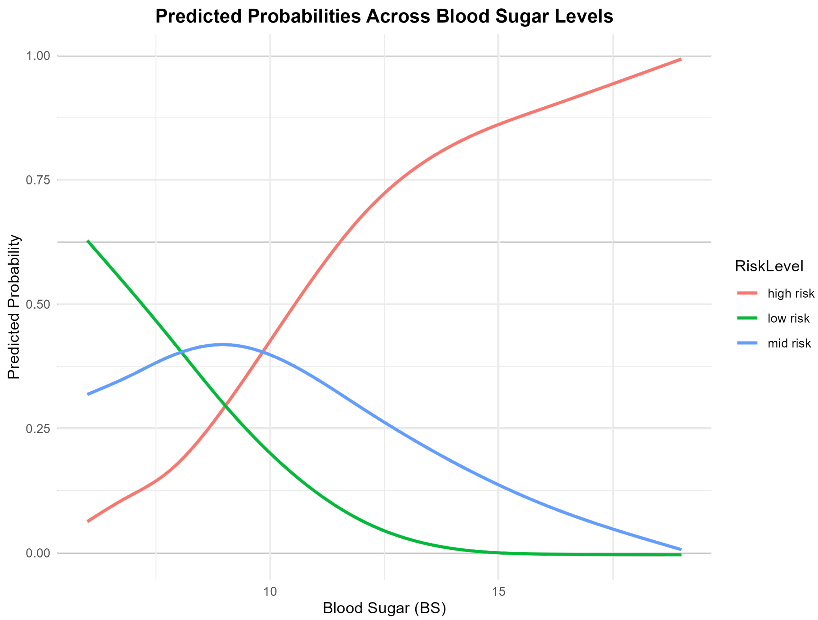


Figure 3: Predicted Probabilities Across Blood Sugar Levels

Table 8: Random Forest Model

```
Call:
 randomForest(formula = RiskLevel ~ ., data = train_data, ntree = 500,      importance = TRU
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 18.1%
Confusion matrix:
          high risk low risk mid risk class.error
high risk       197       10       11  0.09633028
low risk          7      265       53  0.18461538
mid risk         20       46      203  0.24535316
```

Model Testing

Now that the model is trained using our train set, we can make predictions on the test set. We find that the random forest model provides a better accuracy score than the multinomial regression model with an accuracy of 0.8910891. Moreover, we can assess feature importances with a random forest model.

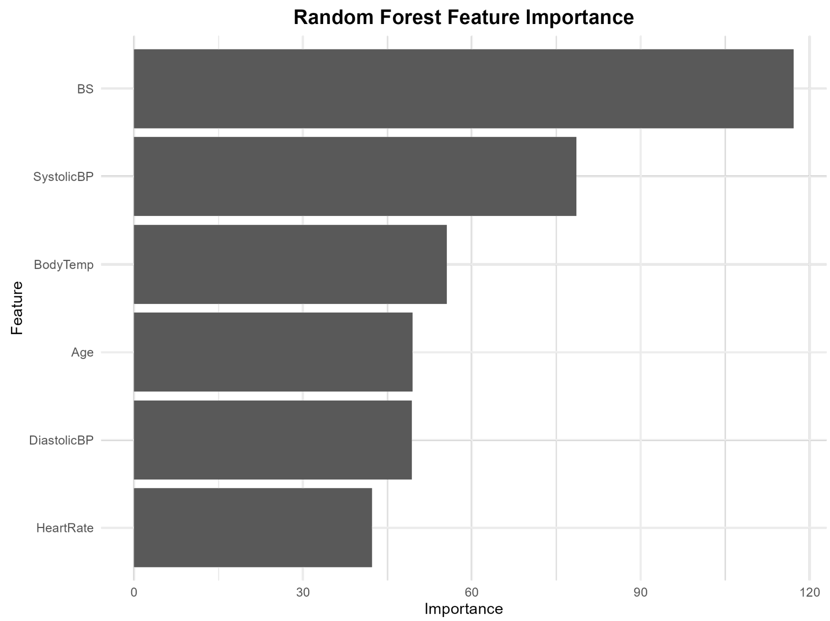


Figure 4: Random Forest Feature Importances

Based on Figure 4 above, we see that the model identifies blood sugar, systolic blood pressure, and body temperature as the top predictors of maternal health risk (i.e., have the most predictive power). With blood sugar specifically, we see that it's feature importance reaches over 100 indicating that it is highly influential in predicting maternal health risk compared to the other features.

Results

Comparison of Results

From Figure 6, random forest has yielded a 0.8910891 accuracy score, which is 0.2772277 higher than the multinomial logistic regression accuracy score and 0.490099 higher than of the baseline model.

Table 9: Baseline Confusion Matrix

```
# A tibble: 9 x 4
  True      Predicted Frequency Percentage
  <chr>     <chr>         <dbl>      <dbl>
1 high risk high risk         0       NA
2 low risk  high risk        54       26.7
3 mid risk  high risk         0       NA
4 high risk low risk          0       NA
5 low risk  low risk         81       40.1
6 mid risk  low risk          0       NA
7 high risk mid risk          0       NA
8 low risk  mid risk         67       33.2
9 mid risk  mid risk          0       NA
```

Table 10: Multinomial Logistic Regression Confusion Matrix

```
# A tibble: 9 x 4
  True      Predicted Frequency Percentage
  <chr>     <chr>         <dbl>      <dbl>
1 high risk high risk        38       76
2 low risk  high risk         2        2.1
3 mid risk  high risk        14       25.5
4 high risk low risk          2        4
5 low risk  low risk         62       63.9
6 mid risk  low risk         17       30.9
7 high risk mid risk         10       20
8 low risk  mid risk         33       34
9 mid risk  mid risk         24       43.6
```

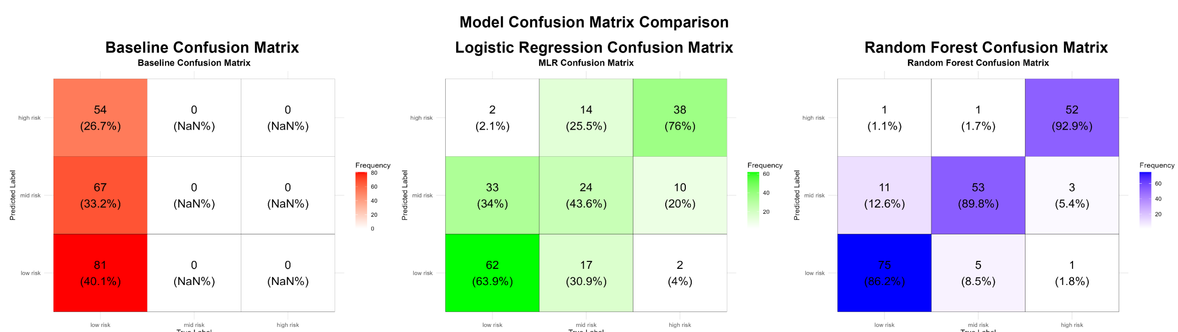


Figure 5: ML Model Confusion Matrices

Table 11: Random Forest Confusion Matrix

```
# A tibble: 9 x 4
  True      Predicted Frequency Percentage
  <chr>     <chr>         <dbl>      <dbl>
1 high risk high risk        52       92.9
2 low risk  high risk         1        1.1
3 mid risk  high risk         1        1.7
4 high risk low risk          1        1.8
5 low risk  low risk         75       86.2
6 mid risk  low risk          5        8.5
7 high risk mid risk          3        5.4
8 low risk  mid risk         11       12.6
9 mid risk  mid risk         53       89.8
```

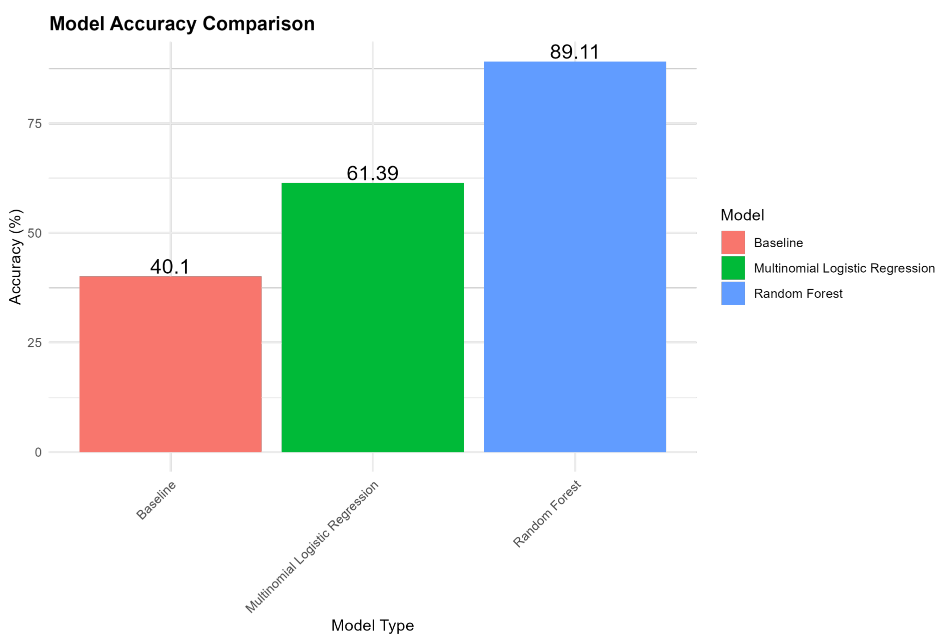


Figure 6: ML Model Accuracies (%)

Discussion

Best performing model

As shown in the bar graph above, our analyses suggest that maternal health risk during pregnancy can be predicted with up to a 89.11% accuracy using our random forests decision trees model. The second best model was the multinomial logistic regression that had a 61.39% accuracy. Both models performed better than the baseline (40.1%). Both models performing better than the baseline is expected as well as the random forest performing better than the logistic regression as past literature have found similar results (Mu, Yan, and Zhu 2023; Ukrit et al. 2024).

Interpretation

The multinomial logistic regression, while less accurate, is more interpretable and gives us an idea of how a 1-unit increase in a variable is associated with a change in the odds of being in a certain risk category. For example, the multinomial logistic regression suggests that a 1 unit increase in Blood Sugar level is associated with an increase in the odds of being in `high risk` compared to `low risk` by a factor of 2.1974809.

The best predictors for `high risk` compared to `low risk` were body temperature (OR = 2.1442957, $p < .001$) and blood sugar (OR = 2.1974809, $p < .001$). A one unit increase in both was associated with a more than double increase in the odds of being in the `high risk` category.

The best predictors for `medium risk` compared to `low risk` were also body temperature (OR = 1.4898473, $p < .001$) and blood sugar (OR = 1.4546199, $p < .001$).

Both the multinomial logistic regression model and the Random Forest model performed best when predicting `high risk`.

Impact

Our analyses show that body temperature and blood sugar are both relatively strongly associated with increasing maternal health risk. We do however acknowledge that our models do not necessarily imply a causal effect such that reducing blood sugar or body temperature will reduce your maternal health risk. Additionally, our models are limited by the number of variables we accounted for. Other factors such as age, parental health conditions, and many more would improve the generalizability of our models.

Our analyses should therefore not be used as guidelines for pregnant mothers. Now that we have have further evidence that blood sugar and body temperature are associated with maternal health risk, future research could explore the potential causal mechanism of these relationships. Future research may also explore whether this effect remains constant across age or whether certain age groups are more susceptible to the effects of blood sugar/body temperature on maternal health risk.

References

Ahmed, Marzia. 2020. "Maternal Health Risk." UCI Machine Learning Repository.

Bajaj, Disha, Ritika Kumari, and Poonam Bansal. 2023. "Risk Level Prediction for Maternal Health Using Machine Learning Algorithms." *2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI)*, November, 405–9. https://doi.org/10.1109/iccsai59793.2023.10421156.

Mu, Chenyu, Zexuan Yan, and Yidi Zhu. 2023. "Prediction of Maternal Health Risk Based on Physiological Indicators." *Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science*, October, 578–84. https://doi.org/10.1145/3644116.3644212.

Ukrit, M.Ferni, R. Beaulah Jeyavathana, Aluru Leela Rani, and Vasa Chandana. 2024. "Maternal Health Risk Prediction with Machine Learning Methods." *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, February, 1–9. https://doi.org/10.1109/ic-etite58242.2024.10493737.

WHO. 2024. "Maternal Mortality Fact Sheet." World Health Organization. https://www.who.int/news-room/fact-sheets/detail/maternal-mortality.

———. 2025. "Maternal Health." World Health Organization. https://www.who.int/health-topics/maternal-health#tab=tab_1.