Predicting Maternal Health Risk Using Classification and Machine Learning

By Victoria Hernandez

Introduction

The dataset that will be analyzed in this report is titled “Maternal Health Risk” and it was created by Marzia Ahmed. Maternal Health is a highly important health issue that affects women worldwide. According to the World Health Organization (WHO), in 2020 around 287,00 women died both during and after childbirth. (World Health Organization & Shimizu, n.d.) Some of the leading causes include infection, high blood pressure, and excessive blood loss. (World Health Organization & Shimizu, n.d.) Predictive models can help analyze different indicators such as heart rate, blood pressure, blood glucose levels, age, and body temperature.

To analyze this data set exploratory, classification, and machine learning methods will be applied. The exploratory methods used include analysis of the mean, minimum, maximum, standard deviation, and quartiles. The preprocessing methods used to prepare for classification are the holdout method and partitioning of the dataset. The machine learning methods used include K-Nearest Neighbor, Naïve Bayes, and Decision Tree. The upcoming sections the report will explain the data, methods used, explain the machine learning models, provide a conclusion of the key findings, and discuss potential future analysis.

Data

The dataset used for this report was sourced from the UC Irvine Machine Learning Repository. The creator of the dataset Marzia Ahmed sourced the data from different hospitals, maternal health care, and community clinics from rural areas of Bangladesh using an IoT based risk monitoring system. (Ahmed, 2020) An IoT-based risk monitoring system uses devices that are interconnected to collect and then analyze data in real time. (Aadil, Khan, Yu, Ali, & Kumar, 2024) The dataset contains 1,013 patient records and each records includes the patients age, blood glucose level (BS), body temperature, systolic and systolic blood pressure, heart rate, and risk level. The target variable for this analysis is the patients Risk Level.

Table 1. Attributes from dataset.

Attributes	Description	Type	Range
Age	Age in years during pregnancy.	Numeric	10-70 years
SystolicBP	Upper value of blood pressure in mmHg.	Numeric	70-160 mmHg
DiastolicBP	Lower value of blood pressure in mmHg.	Numeric	49-100 mmHg
BS	Blood glucose levels in terms of molar concentration.	Numeric	6-19 mmol/L
BodyTemp	Body temperature in Fahrenheit.	Numeric	98-103 Fahrenheit

HeartRate	Normal resting heart rate in bpm.	Numeric	7-90 bpm
RiskLevel	Predicated risk intensity during pregnancy.	Categorical	Low Risk Mid Risk High Risk

Methods

Exploratory Data Analysis:

- The descriptive statistics used in this analysis are metrics such as the mean, minimum, maximum, standard deviation, and quartiles.

Data preprocessing:

- The variable RiskLevel was converted into a factor for classification in R.
- The dataset was partitioned by 70% for training and 30% for testing using the holdout method.

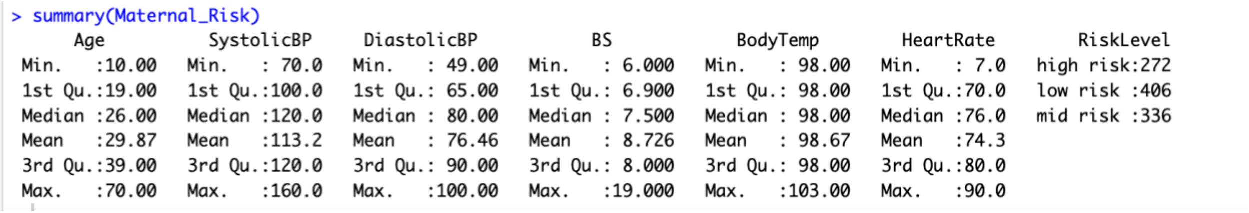
Machine Learning models:

- K-Nearest Neighbors: Apobalistic supervised machine learning algorithm that uses proximity to make predictions about data grouping. (IBM, n.d.)
- Naïve Bayes: A supervised machine learning algorithm that uses Bayes Theroem used for classification. (IBM, n.d.)
- Decision Tree: A tree structure where each node represents an attribute test, each branch represents an outcome, and a leaf has a class label. (McGuire, 2025)
- The entire training set was used to train each algorithm.

Results

Exploratory statistics:

Figure 1. Exploratory statistics of the Maternal health Risk for the patient data.



```
> summary(Maternal_Risk)
      Age          SystolicBP      DiastolicBP          BS            BodyTemp        HeartRate        RiskLevel
 Min.   :10.00   Min.   : 70.0   Min.   : 49.00   Min.   : 6.000   Min.   : 98.00   Min.   : 7.0   high risk:272
 1st Qu.:19.00   1st Qu.:100.0   1st Qu.: 65.00   1st Qu.: 6.900   1st Qu.: 98.00   1st Qu.:70.0   low risk :406
 Median :26.00   Median :120.0   Median : 80.00   Median : 7.500   Median : 98.00   Median :76.0   mid risk :336
 Mean   :29.87   Mean   :113.2   Mean   : 76.46   Mean   : 8.726   Mean   : 98.67   Mean   :74.3
 3rd Qu.:39.00   3rd Qu.:120.0   3rd Qu.: 90.00   3rd Qu.: 8.000   3rd Qu.: 98.00   3rd Qu.:80.0
 Max.   :70.00   Max.   :160.0   Max.   :100.00   Max.   :19.000   Max.   :103.00   Max.   :90.0
```

Age: Based on the 1st and 3rd quartiles most of the recorded ages are from 19 to 39 years.

Heart Rate: The average heart rate was 74.3 bpm based on the mean.

Body Temperature: The average body temperature is 98.67 F, which is in the normal range for body temperature.

Blood Glucose(BS): The mean is 8.73 mmol/L and the median is 7.5 mmol/L meaning it is skewed right.

Table 1. K-Nearest Neighbors Confusion Matrix:

Actual class/Predicted class	High Risk	Low Risk	Mid Risk	Total
High Risk	59	3	11	73
Low Risk	8	81	35	124
Mid Risk	14	37	54	105

- KNN predicted High Risk accurately with 59/73 or 81%
- Mid Risk had misclassified predictions with 37/105 as Low Risk and 14/105 as High Risk.
- Low Risk had 35/124 misclassified as Mid Risk.

Tabel 2. K-Nearest Neighbors Statistics:

Statistic	High Risk	Low Risk	Mid Risk
Accuracy	0.8325	0.7159	0.6438
Precision	0.8082	0.6532	0.5143
Recall	0.7284	0.6612	0.5400
F1	0.7662	0.6612	0.5268

- Out of all three models KNN had the most balanced overall results.
- KNN had the highest recall for Mid Risk meaning it could be better at detecting Mid Risk cases.

Table 3. Naïve Bayes Confusion Matrix:

Actual class/Predicted class	High Risk	Low Risk	Mid Risk	Total
High Risk	62	0	14	76
Low Risk	8	105	51	164
Mid Risk	11	16	35	62

- Naïve Bayes had no incorrect predictions for High Risk as Low Risk. However, there were 14/76 misclassified as Mid Risk.
- Low Risk was the best predicted by the model with 105/164 or 64% correctly predicted. However, there were 51/164 or ~0.31% predicted as Mid Risk.
- Mid Risk was predicted well by the model with about 43% predicted as either Low or High Risk.

Table 4. Naïve Bayes Statistics:

Statistic	High Risk	Low Risk	Mid Risk
Accuracy	0.8510	0.7709	0.6082
Precision	0.8158	0.6402	0.5645
Recall	0.7654	0.8678	0.3500
F1	0.7898	0.7368	0.4321

- Naïve Bayes had the best recall of 0.8678 and F1 0.7368 for Low Risk.
- Mid Risk shows weak scores meaning the model may not be able to detect these cases as wells the other models.

Table 5. Decision Tree Confusion Matrix:

Actual class/Predicted class	High Risk	Low Risk	Mid Risk	Total
High Risk	72	2	21	95
Low Risk	6	61	26	93
Mid Risk	3	58	53	114

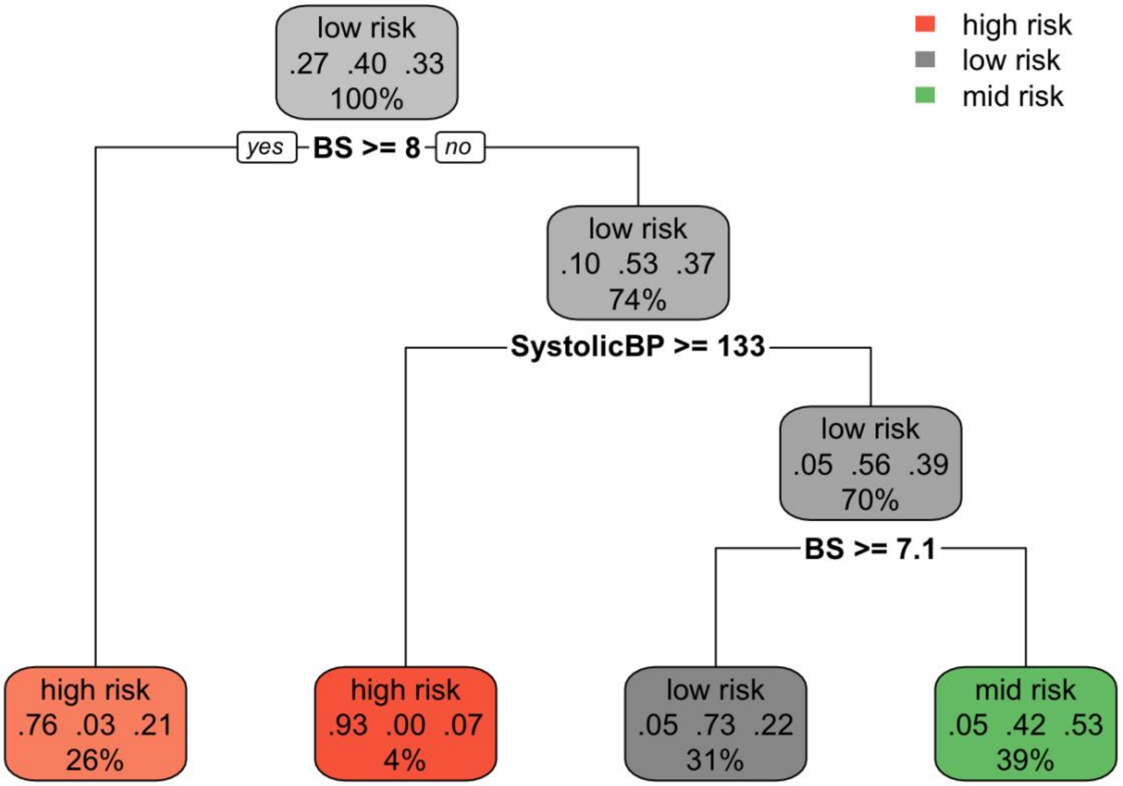
- The decision tree classified the High Risk well with 72/95 or ~76% and only 2 Low Risk misclassifications.
- The tree had difficulty classifying Low Risk at 61/93 or ~66% and misclassified as Mid Risk at 26/93 or ~28%.
- The tree also had a hard time classifying Mid Risk correctly with 58/114 or ~51% classified as Low Risk.

Table 6. Decision Tree Statistics:

Statistic	High Risk	Low Risk	Mid Risk
Accuracy	0.8924	0.6637	0.6140
Precision	0.7579	0.6559	0.4649
Recall	0.8889	0.5041	0.5300
F1	0.8182	0.5701	0.4953

- The decision tree had the best High Risk recall of 0.8889 and the best F1 of 0.8182.
- The decision tree had weak Low Risk scores especially Low Risk’s Recall at 0.5041.

Figure 2. Decision Tree Visualization



Overall analysis: The decision tree was the most ideal for high risk case detection. Naïve Bayes was the most ideal for classifying low risk cases detection. K-nearest neighbors was had the best overall performance for all classes and metrics especially for mid risk having the highest recall score compared to the other models, where the other two models struggled. However, it seems like al the models did have trouble predicting mid class likely because there were fewer instances in the training set compared to the other classes.

Conclusion and Discussion

This report used K-Nearest Neighbors, Naïve Bayes, and a Decision Tree to classify maternal health risk using data from pregnant patients. The data was partition using the holdout method splitting the data 70/30 and the accuracy, recall, precision, and F1 score was collected and analyzed.

Key Finding:

- K-Nearest Neighbors performed the best for all three classes and had the most balanced results. It performed the best on the Mid Risk class with an F1 Score of 0.5268.
- Naïve Bayes performed the best with the Low Risk class with an F1 score of 0.7368 and a recall of 0.8678.
- The Decision Tree performed the best with the High Risk class with the highest recall of 0.8889 and F1 score of 0.8182.

Limitation:

- The dataset only uses a limited set of biological features of each patient that may not show all the possible contributors of maternal health risk. There could be other unreported factor that could be contributed that may be affecting the data.
- Since the data was taken from a Bangladesh there could be specific regional contributors that could be affecting the results.

Potential future analysis:

Future analysis could include using cross validation to improve the machine learning models. Collecting more biological information about the patients could also provide future insights what cause a patient to be categorized as a high risk for maternal health complications.

References

Ahmed, M. (2020). Maternal Health Risk. [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DP5D.

World Health Organization. (n.d.). Maternal health. World Health Organization. Retrieved March 30, 2025, from https://www.who.int/health-topics/maternal-health#tab=tab_1

IBM. (n.d.). Naive Bayes. IBM. Retrieved March 30, 2025, from https://www.ibm.com/think/topics/naive-bayes

Aadil, M., Khan, S. A. R., Yu, Z., Ali, S. M., & Kumar, A. (2024). Analysis of Internet of Things implementation barriers in the cold supply chain: An integrated ISM-MICMAC and DEMATEL approach. ResearchGate. https://www.researchgate.net/publication/382209123

McGuire, M. (2025). Decision Trees & Naïve Bayes [Slides 1-31]. CIS 468: Data Science, Towson University.

McGuire, M. (2025). Supervised Learning and Classification [Slides 1-52]. CIS 468: Data Science, Towson University.