Group Members: Emmanuel Adeoti, Carolyn Bollerman, Hilary Henshaw
15 December 2024
DSP 556 Final Project

Predicting Maternal Survival: A Machine Learning Approach to Health Risk Analysis

Introduction

Maternal health, encompassing the physical and mental well-being of women during pregnancy, childbirth, and the postpartum period, is an ongoing public health crisis. It profoundly impacts mothers, families and communities throughout the world. Despite significant advancements in healthcare, maternal mortality remains a global public health crisis, particularly in resource-constrained regions where access to timely medical care is limited. Understanding and addressing the factors that contribute to maternal health risks is essential for developing effective interventions and ultimately reduce preventable deaths.

This study investigates a critical research question: How can health factors be used to predict maternal survival during pregnancy? To answer this, we leverage data-driven approaches to analyze patterns in health indicators and predict maternal health risks using machine learning. Our research utilizes the Maternal Health Risk dataset from the UCI Machine Learning Repository, which comprises 1,014 observations and seven attributes, including key health indicators such as systolic blood pressure, diastolic blood pressure, blood sugar levels, and heart rate. These attributes are used to classify maternal health risks into three categories: low, medium, and high. To enhance the study's real-world applicability, we also incorporate data from IoT-enabled wearable devices monitoring pregnant women in Bangladesh. These devices record physiological metrics in real time, providing critical health alerts to mothers and their families in emergency situations.

By integrating advanced machine learning models, we aim to develop accurate predictive algorithms for maternal health risk classification. The dataset is structured for classification tasks, making it suitable for exploring how physiological and demographic factors contribute to maternal health risks. The methodology employed in this study involved data preprocessing to clean the dataset and handle missing values by imputing with column means, an exploratory data analysis to understand feature distributions, relationships, and correlations between predictors. Additionally, supervised learning techniques were deployed including logistic regression, decision trees, random forests, and support vector machines to classify maternal health risks. To ensure robustness of outputs, an evaluation of model performance was conducted using metrics such as accuracy, F-1 score, and area under the receiver operating characteristic curve (AUC), with stratified cross-validation. Finally, a feature importance analysis was conducted using tree-based models to identify the most influential predictors of maternal health risks. A neural network model was also implemented to explore its performance in classifying maternal health risks. The model architecture included multiple dense layers with ReLU activation and a softmax output layer for multiclass classification. Its performance was evaluated through cross-validation and testing.

This research builds on existing literature that explores the integration of IoT technology and machine learning to address maternal health challenges. A notable study by Ahmed et al. (2020) examined the use of IoT-enabled wearable devices for real-time monitoring of pregnant women in rural areas, coupled with machine learning algorithms for risk classification. Their findings demonstrated the potential of combining predictive analytics with IoT technology to

improve maternal health outcomes. Inspired by this work, our study extends the scope by comparing predictions from the UCI dataset with real-world IoT data to validate the reliability of the proposed methods across diverse settings.

Through this collaborative project, we aim to contribute to the global effort to reduce maternal mortality by providing actionable insights and developing scalable solutions. The findings of this study have the potential to inform both clinical practices and policy decisions, particularly in resource-constrained settings in which maternal death rates are extremely prevalent. By leveraging machine learning models, this study demonstrates how data-driven approaches can be applied to maternal health risk classification. Although the analysis is limited to the UCI dataset, it provides a generalizable foundation for developing predictive tools that could guide clinical decision-making and improve maternal health outcomes. This paper will detail the dataset, methodology, and key findings of our research, demonstrating the transformative potential of integrating machine learning with real-world health data.

Exploratory Data Analysis[1]

The Maternal Health Risk dataset, sourced from the UCI Machine Learning Repository[2], consists of 1,014 observations with seven attributes: Age, Systolic Blood Pressure (SystolicBP), Diastolic Blood Pressure (DiastolicBP), Blood Sugar (BS), Body Temperature (BodyTemp), Heart Rate, and Risk Level (Low, Mid, High). These variables provide a comprehensive overview of maternal health, capturing key physiological and demographic factors that influence maternal risk.

The age of pregnant women in the dataset ranges from 10 to 70 years, with an average age of 29.87 years and a standard deviation of 13.47 years. The interquartile range (IQR) indicates that 50% of the participants are between 19 and 39 years old, with a median age of 26 years. Systolic blood pressure (SystolicBP) values range from 70 to 160 mmHg, with a mean of 113.20 mmHg and a standard deviation of 18.40 mmHg. The median systolic pressure is 120 mmHg, and most values fall between 100 and 120 mmHg, indicating that many of the participants have blood pressure within a normal range. Similarly, diastolic blood pressure (DiastolicBP) values range from 49 to 100 mmHg, with a mean of 76.46 mmHg and a standard deviation of 13.89 mmHg. The median diastolic pressure is 80 mmHg, with half of the values lying between 65 and 90 mmHg.

Blood sugar levels (BS) exhibit a wide range, from 6.0 to 19.0 mmol/L, with a mean of 8.73 mmol/L and a standard deviation of 3.29 mmol/L. The median blood sugar level is 7.5 mmol/L, with the IQR showing that 50% of observations lie between 6.9 and 8.0 mmol/L. Body temperature (BodyTemp) measurements range from 98.0°F to 103.0°F, with a mean of 98.67°F and a standard deviation of 1.37°F. Most participants fall within the normal temperature range of 98.0°F to 98.6°F. Heart rates in the dataset range from an anomalously low value of 7 beats per minute (bpm) to 90 bpm, with a mean of 74.30 bpm and a standard deviation of 8.09 bpm. The median heart rate is 76 bpm, and half of the participants have heart rates between 70 bpm and 80 bpm.

The target variable, RiskLevel, categorizes maternal health risks into three classes: Low Risk, Mid Risk, and High Risk. The dataset's balanced distribution of risk levels[3] was analyzed

[1] See appendix
[2] UCI Machine Learning Repository, "Maternal Health Risk Data Set," accessed December 14, 2024, https://archive.ics.uci.edu/dataset/863/maternal+health+risk.
[3] See appendix

in the exploratory data analysis phase, ensuring suitability for classification tasks. These descriptive statistics provide key insights into the dataset's structure and variability, highlighting critical predictors such as age, blood pressure, and blood sugar levels, which are essential for modeling maternal health risks.

Methodology

The Maternal Health Risk was inspected for missing values using the isnull() method. Since the dataset contained no missing values, imputation or removal was not necessary. This ensured that all 1,014 entries were retained for the analysis. The target variable, RiskLevel, was encoded into numeric categories to facilitate machine learning analysis. Low risk was encoded to 0, mid risk was encoded to 1, and high risk was encoded to 2. Data splitting and scaling was also performed. The dataset was split into training and testing subsets using a 70-30 stratified split to ensure proportional representation of the RiskLevel classes in both subsets. The features were standardized using a StandardScaler, transforming them to have a mean of 0 and a standard deviation of 1. This step ensured that models sensitive to feature magnitudes, such as Logistic Regression and Support Vector Machines (SVM), could perform optimally.

To predict maternal health risk, four supervised learning algorithms were selected for classification: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Support Vector Machine. These models were chosen to provide a balanced exploration of both linear and non-linear approaches, as well as the capability to assess feature importance. In this study, Logistic Regression served as a baseline model to predict maternal health risks based on linear relationships between features such as age, blood pressure, and blood sugar. Its simplicity and interpretability make it a valuable starting point, providing a benchmark to evaluate the performance of more complex models. The Decision Tree Classifier was chosen to capture non-linear relationships in the data, which are common in health-related variables. Decision Trees also offer interpretability by visually demonstrating which features (e.g., age, blood sugar) and thresholds are most critical for classifying maternal health risks. However, due to their propensity to overfit, this model's performance was compared to ensemble methods like Random Forests to assess robustness. By building multiple trees and averaging their predictions, Random Forests improved classification accuracy and reduced overfitting. For this project, Random Forests also provided valuable feature importance scores, which highlighted the predictors most strongly associated with maternal health risks. This helped identify key contributors, such as systolic blood pressure and heart rate, in determining risk levels. Support Vector Machines were included to leverage their ability to handle complex, non-linear relationships. In this project, an RBF (radial basis function) kernel was applied to capture intricate patterns in the physiological and demographic data.

Each algorithm was trained on the standardized training dataset and evaluated on the test dataset. To ensure the models were robust and their performance generalized well to unseen data, five-fold stratified cross-validation was employed. This approach ensured that each RiskLevel category (Low, Medium, High) was proportionally represented in both training and validation subsets during cross-validation. The models were then evaluated using accuracy, F-1 score, and AUC.

By applying these models, the study was able to compare the effectiveness of linear and non-linear algorithms, assess feature importance, and identify the most suitable approach for predicting maternal health risks. The insights gained from these supervised learning methods

directly informed the development of actionable tools for health risk assessment, with potential applications in clinical and real-world settings.

Main Findings

This study sought to predict maternal health risks using supervised learning models, focusing on three classes: Low Risk (0), Medium Risk (1), and High Risk (2). The primary models evaluated include stacking classifiers with Linear Discriminant Analysis (LDA) and Support Vector Classifier (SVC) as final estimators, as well as a Decision Tree Classifier. The results demonstrate varying levels of success across the models, particularly in handling class imbalances and achieving robust classification performance.

Stacking Classifier with LDA as Final Estimator[4]

The stacking classifier with LDA as the final estimator achieved a strong overall performance, correctly classifying 82.3% of instances in the test set. This model demonstrated a high degree of accuracy in predicting classes 0 (Low Risk) and 2 (High Risk). However, there is room for improvement in predicting class 1 (Mid Risk), as its performance in this category was comparatively weaker. The challenges in classifying class 1 may be due to its intermediate nature, where feature values overlap significantly with the other two classes.

Stacking Classifier with SVC as Final Estimator[5]

The stacking classifier with SVC as the final estimator exhibited slightly better overall performance than the LDA-based model, with 83% of instances in the test set correctly classified. Similarly, the LDA-based stacking model classifier demonstrated strong performance in predicting classes 0 (Low Risk) and 2 (High Risk) but struggled with class 1 (Mid Risk). The SVC's ability to model non-linear relationships likely contributed to its improved classification accuracy compared to the LDA-based model.

Decision Tree Classifier[6]

The Decision Tree Classifier outperformed both stacking classifiers in terms of overall test accuracy, correctly classifying 84% of instances in the test set. It demonstrated the highest true positive (TP) rate for class 1 (Mid Risk), which is a notable strength compared to the stacking models. The result highlights the Decision Tree's capability in effectively capturing the distinct thresholds and non-linear relationships within the dataset. Furthermore, it also performed well in predicting Low Risk and High Risk, which makes it the most balanced model among those evaluated.

All three models demonstrated a strong performance in predicting Low and High Risk cases, underscoring the ability of machine learning to identify maternal health outcomes. For example, extreme values in systolic blood pressure or blood sugar are likely driving accurate classifications. This aligns with existing maternal health literature which emphasizes the

[4] See appendix
[5] See appendix
[6] See appendix

importance of monitoring these indicators to prevent complications such as preeclampsia or gestational diabetes. An early indication of High Risk cases can lead to timely intervention, which in turn could potentially reduce maternal mortality rates.

Discussion and Conclusion

This study demonstrates the potential of machine learning to enhance maternal health care through the accurate prediction of maternal risk levels. By leveraging physiological and demographic features such as systolic blood pressure, blood sugar, and heart rate, the models achieved strong performance in identifying Low and High Risk cases. The findings underscore the promise of predictive analytics in supporting clinical decision-making, proactive monitoring, and targeted interventions for at-risk populations.

The results of this study align with global maternal health priorities by providing a scalable, data-driven framework to identify risk categories and guide timely interventions. One of the most significant outcomes of this study is the ability to classify Low and High Risk cases with a high degree of accuracy, which demonstrates the reliability of key health metrics in predicting maternal outcomes. These findings affirm the value of machine learning models as complementary tools in maternal health care, where early identification of possible complications can prevent adverse outcomes and save lives. Additionally, this study underscores the potential for integrating predictive models with real-time monitoring systems, such as the IoT enabled wearable devices, to enhance maternal health care delivery. This is especially helpful in resource-constrained settings.

The study also further highlighted the challenge in classifying Medium Risk cases. This speaks to the critical gap in current medical approaches, in which many patients in that category often have overlapping and confounding features with the other two classes, leading to higher misclassification rates. Addressing this gap will require more data and a more nuanced approach to modeling, such as a longitudinal analysis or the inclusion of contextual variables like socioeconomic status, pre-existing conditions, and access to prenatal care.

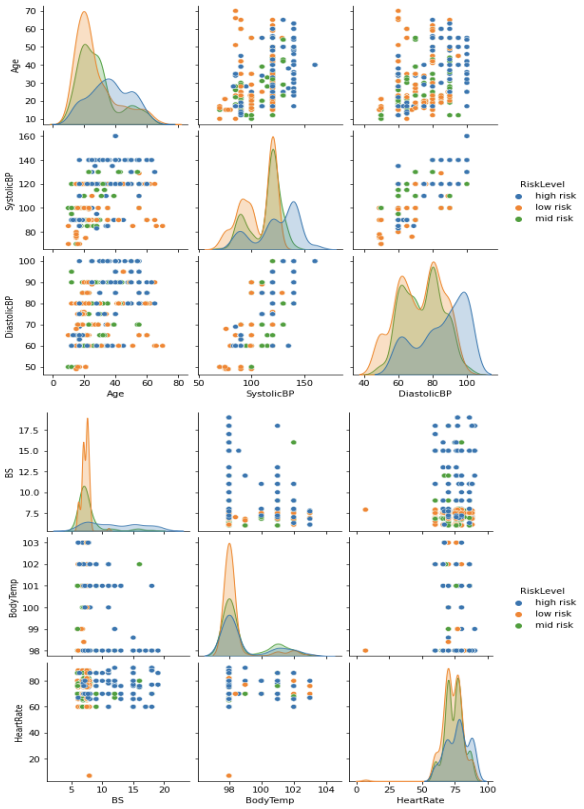
This project highlights the potential of machine learning to improve maternal health outcomes by enabling accurate risk predictions and identifying key health indicators. While the models performed well overall, the difficulty in predicting Medium Risk cases underscores the need for further research and innovation to address this complexity. By integrating richer datasets, contextual variables, and advanced modeling techniques, future efforts can refine predictive analytics in maternal health care.

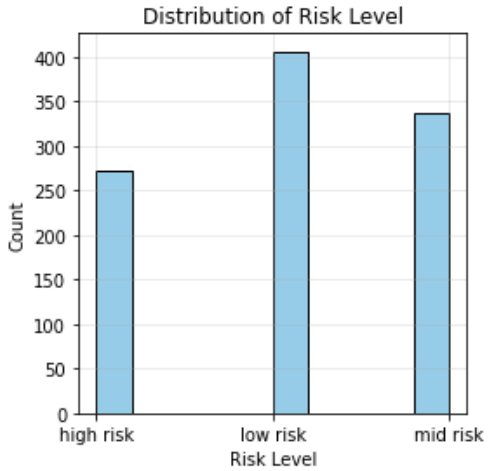
Finally, this study contributes to the broader goal of reducing maternal mortality by offering a scalable, interpretable, and actionable approach to risk assessment. It also serves as a foundation for future research aimed at addressing the nuanced challenges of maternal health, paving the way for data-driven, patient-centered care.

Appendix

1. Descriptive Statistics

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1014 entries, 0 to 1013
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Age          1014 non-null   int64
 1   SystolicBP   1014 non-null   int64
 2   DiastolicBP  1014 non-null   int64
 3   BS           1014 non-null   float64
 4   BodyTemp     1014 non-null   float64
 5   HeartRate    1014 non-null   int64
 6   RiskLevel    1014 non-null   object
dtypes: float64(2), int64(4), object(1)
memory usage: 55.6+ KB
None


First 5 Rows of the Dataset:
   Age  SystolicBP  DiastolicBP    BS  BodyTemp  HeartRate  RiskLevel
0   25         130           80  15.0      98.0         86  high risk
1   35         140           90  13.0      98.0         70  high risk
2   29          90           70   8.0     100.0         80  high risk
3   30         140           85   7.0      98.0         70  high risk
4   35         120           60   6.1      98.0         76   low risk


Descriptive Statistics of Numerical Columns:
               Age   SystolicBP  DiastolicBP           BS     BodyTemp  \
count  1014.000000  1014.000000  1014.000000  1014.000000  1014.000000
mean     29.871795   113.198225    76.460552     8.725986    98.665089
std      13.474386    18.403913    13.885796     3.293532     1.371384
min      10.000000    70.000000    49.000000     6.000000    98.000000
25%      19.000000   100.000000    65.000000     6.900000    98.000000
50%      26.000000   120.000000    80.000000     7.500000    98.000000
75%      39.000000   120.000000    90.000000     8.000000    98.000000
max      70.000000   160.000000   100.000000    19.000000   103.000000

         HeartRate
count  1014.000000
mean     74.301775
std       8.088702
min       7.000000
25%      70.000000
50%      76.000000
75%      80.000000
max      90.000000


Range of Age: Minimum = 10, Maximum = 70


Range of Systolic Blood Pressure: Minimum = 70 mmHg, Maximum = 160 mmHg


Range of Diastolic Blood Pressure: Minimum = 49 mmHg, Maximum = 100 mmHg
```





4 Stacking Classifier with LDA as Final Estimator

```
Confusion Matrix using predict:
[[69 12  0]
 [13 51  3]
 [ 1  3 51]]
Confusion Matrix assuming perfection:
[[81  0  0]
 [ 0 67  0]
 [ 0  0 55]]
```

5 Stacking Classifier with QDA as Final Estimator

```
Confusion Matrix using predict:
[[64 17  0]
 [ 4 57  6]
 [ 1  4 50]]
Confusion Matrix assuming perfection:
[[81  0  0]
 [ 0 67  0]
 [ 0  0 55]]
```

6 Decision Tree Classifier

```
Confusion Matrix using predict:
[[69 12  0]
 [15 51  1]
 [ 6  4 45]]
Confusion Matrix assuming perfection:
[[81  0  0]
 [ 0 67  0]
 [ 0  0 55]]
```

Works Cited

Ahmed, Marzia and Mohammod Abul 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)* (2020): 1-6.

Ahmed, Marzia. "Maternal Health Risk." UCI Machine Learning Repository, 2020, https://doi.org/10.24432/C5DP5D.