Final Report: Maternal Health Risk

Submitted to:

Prof. David Hubbard

School of Computing and Digital Media

DePaul University

Report Prepared By:

Natalia Guzman

March 22, 2025

1. Introduction

Maternal health is an important topic in public health that encompasses the well-being of pregnant women during the prenatal, delivery, and postpartum stages. Different complications such as hypertension, gestational diabetes, or other metabolic problems, can endanger both the mother and the child even more in low-income countries and environments. Therefore, the early identification of these factors is primordial to reduce the likelihood of adverse outcomes and to provide timely interventions. Additionally, by studying this problem through data, we can help build better protocols for early diagnosis and treatment.

The main goal of this project is to develop a predictive model that classifies mothers into low, mid, and high-risk by studying features such as age, blood sugar, and blood pressure. By exploring different machine learning models like decision trees, logistic regression, random forests, gradient boosting, and ensemble methods, I aim to identify the best predictive approach and to understand how effective machine learning can help enhance data-based clinical decision-making.

2. Data Description

For my project, I used a dataset from the UC Irving Machine Learning Repository. It contains information about 1013 individuals and it has 6 features. The 6 features in the dataset are numerical and they provide information about women's age, systolic blood pressure, diastolic blood pressure, blood glucose in mmol/L, body temperature in Fahrenheit, and heart rate. The target variable is the risk level which ranks from low to high.

3. Data Preprocessing and Data Analysis

For the data preprocessing, I decided to check for missing values and I had none. Then I looked at the data description, where I noticed there were some extremely low values for the heart rate feature which might be outliers. After this, I decided to analyze the distribution of each of the features using countplots. When we look at the target distribution

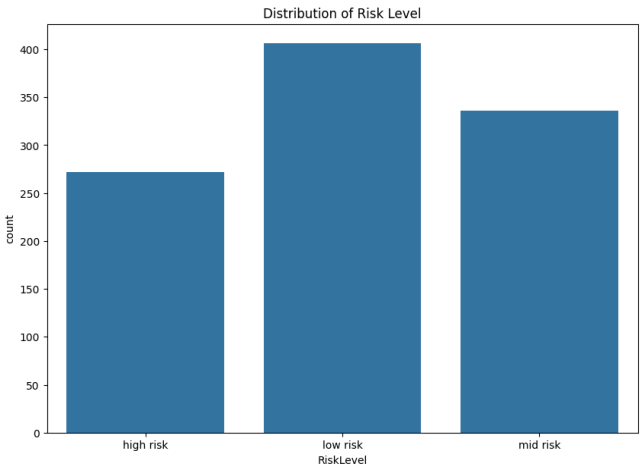

Figure 1. Countplot of Risk Level

we see that we have a higher frequency of 'low risk' individuals, followed by 'mid risk', and 'high risk'. However, we have in general a balanced distribution across the risk levels so we

do not need to worry about working with imbalanced data. When we look at the distribution of the numerical features (Figure 2), we see that regarding age, there is a peak around late

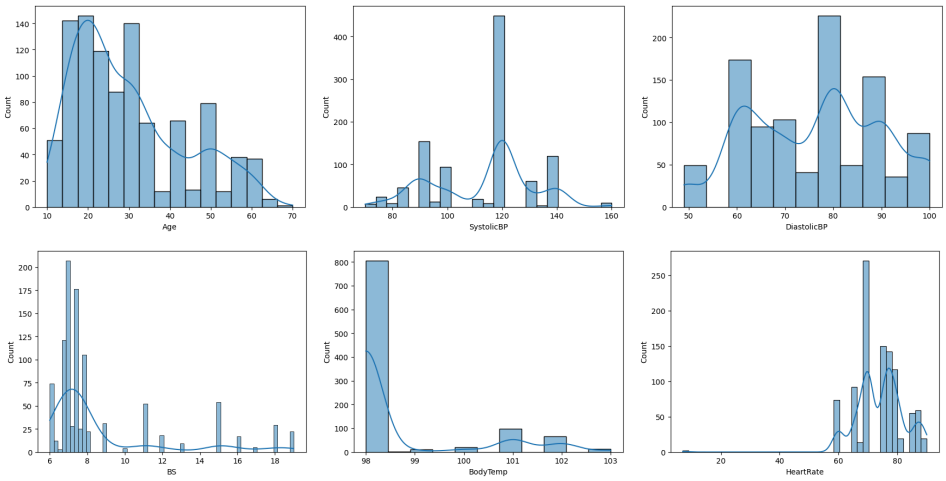

Figure 2. Distribution of Numerical Features

teens and early twenties, and then the count gradually decreases making the graph right-skewed. For systolic blood pressure, we see a peak in 120 which is considered normal and then we see some other values towards the extremes which could mean hypotension or hypertension. For diastolic blood pressure, we see that the most common value is 80, but there are different peaks across the graph demonstrating the variability of this feature. For blood sugar, we have a right-skewed pattern with most of the values being between 6-8. For body temperature, we have a similar pattern with most values located in 98 which is a normal temperature, and then less common values to the higher end. Finally, we see a normal distribution for heart rate, except for an outlier on the lower end of the graph.

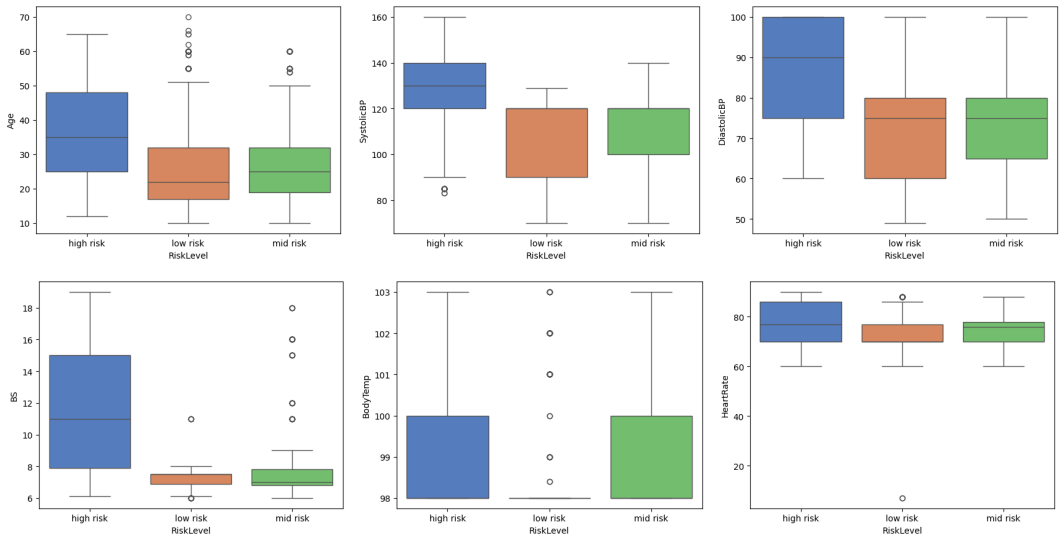

Figure 3. Features vs. Target

Finally, I decided to use boxplots to illustrate the relationship between the features and the target. First, we can see that as age increases, there is a higher tendency towards high risk compared to how the median for low and mid risk are between 20-30 years. For systolic and diastolic blood pressure, we can see that as the pressure is higher, the risk is higher as well, indicating that there is a positive correlation between high blood pressure and maternal risk.

For blood sugar levels, we can also see that higher levels correlate with a higher risk, except for a couple of outliers we see in low and mid-risk. For body temperature, we see that low-risk individuals tend to have normal body temperatures, but for mid and high risk we see more elevated temperatures. For heart rate, we see that higher values appear more commonly in high-risk, but the difference with low and mid-risk is not as notorious or significant as with other features.

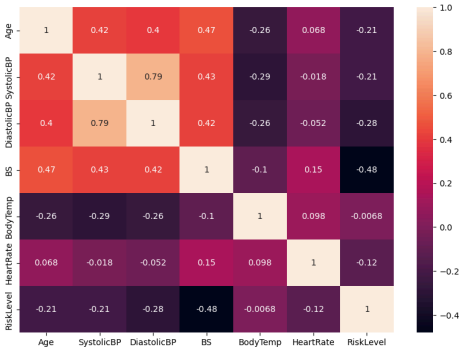

Figure 4. Heatmap of all the variables

After performing the data analysis, I dropped the outliers in the heart rate feature because is not possible for someone to have heart rate values of 7 rpm. Then I performed label encoding for the target which ended up with high risk represented by 0, low risk by 1, and mid-risk by 2. Finally, I used a heatmap to visualize again the correlation between variables and as we can see, there is now a moderate negative correlation between blood sugar and risk level, and a weak negative correlation between systolic blood pressure, diastolic blood pressure, age, and risk level (keeping in mind that now high risk is the lowest number this translates to higher values in these variables, increasing the risk). Finally, I defined X and y, split these into training and testing sets, and performed MinMax scaling to prepare the data to build the models.

4. Model Building and Results

For all the models I chose, I performed a grid search with cross-validation to find the best combination of parameters and validate the model. Then I used the model with the best parameters to make predictions and evaluated the performance using the metric accuracy, precision, and recall.

4.1 Decision Tree

DecisionTree Classification Report:

	precision	recall	f1-score	support
0	0.91	0.88	0.90	104
1	0.88	0.86	0.87	143
2	0.78	0.83	0.81	108
accuracy			0.86	355
macro avg	0.86	0.86	0.86	355
weighted avg	0.86	0.86	0.86	355

Figure 5. Classification Report Decision Tree

For this model, I got that the best parameters were ‘entropy’, max depth none, and minimum samples 2. As seen in Figure 5, the model has an overall accuracy of 86% and it does a good work classifying high-risk individuals with a precision of 91% and a recall of 88%. However, the precision to identify low risk is lower with 88% and with only 78% for mid-risk. This shows that the decision tree model struggles to predict mid-risk individuals.

4.2 Logistic Regression

LogisticRegression Classification Report:

	precision	recall	f1-score	support
0	0.81	0.61	0.69	104
1	0.69	0.87	0.77	143
2	0.45	0.40	0.42	108
accuracy			0.65	355
macro avg	0.65	0.62	0.63	355
weighted avg	0.65	0.65	0.64	355

Figure 6. Classification Report Logistic Regression

For logistic regression, I got that the best parameters were ‘C’ 1 and l2 for the penalty. As a result (Figure 6), the model had an overall accuracy of only 65% and it did not perform as well as the previous model when classifying the different risk levels. For high-risk, it has a precision of 81% but only a recall of 61%, while for the low-risk, the precision was 69%, and for mid-risk 45%. In other words, the model does moderately well predicting low-risk, but the performance is weaker for the other two levels.

4.3 Gradient Boosting

GradientBoosting Classification Report:

	precision	recall	f1-score	support
0	0.91	0.86	0.88	104
1	0.91	0.89	0.90	143
2	0.80	0.87	0.84	108
accuracy			0.87	355
macro avg	0.87	0.87	0.87	355
weighted avg	0.88	0.87	0.87	355

Figure 7. Classification Report Gradient Boosting

For this model, through grid search, I got that the best parameters were a learning rate of 0.1, a max depth of 7, and 100 estimators. The results show that gradient boosting had an accuracy of 87%, which is higher than the previous two models. Additionally, it classifies well the three risk levels with a high precision for high and low-risk (91%) and a good precision for mid-risk (80%). We could say that this model has also a good balance between precision and recall for the three levels, which makes it a robust model.

4.4 Random Forest

RandomForest Classification Report:

	precision	recall	f1-score	support
0	0.91	0.87	0.89	104
1	0.89	0.87	0.88	143
2	0.81	0.87	0.84	108
accuracy			0.87	355
macro avg	0.87	0.87	0.87	355
weighted avg	0.87	0.87	0.87	355

Figure 8. Classification Report Random Forest

For random forest, I got that the best parameters were the criterion ‘gini’, max depth none, and 50 estimators. As a result, the model had an accuracy of 87%, which is the same as the gradient boosting model. However, the performance of random forest was slightly lower for low-risk with a precision of 89% compared to gradient boosting. The precision and recall for high-risk and mid-risk remained almost the same with values of 91% and 87%, and 81% and 87%, respectively.

4.5 Ensemble Model

Ensemble Model Classification Report:

	precision	recall	f1-score	support
0	0.91	0.87	0.89	104
1	0.91	0.89	0.90	143
2	0.81	0.87	0.84	108
accuracy			0.88	355
macro avg	0.88	0.87	0.87	355
weighted avg	0.88	0.88	0.88	355

Figure 9. Classification Report Ensemble Model

For the ensemble model, I initially tried to use the decision tree, random forest, and gradient boosting models. However, the final model had a lower performance than the three of them. Therefore, I decided to create an ensemble model with only the gradient boosting and random forest models using the voting classifier method. The results show a model with a high accuracy of 88%. Additionally, the performance preserves the best of both models. As we can see, the performance for high and low risk is as good as the one we got from the gradient boosting model, and the performance for mid-risk is the same as the random forest model, which was also the highest across all the models for this level.

5. Conclusions

This analysis demonstrates the ability and potential of machine learning to predict maternal health risk based on different physiological factors. By applying four different models and refining them through grid search, I was able to optimize them and find the best parameters for each approach. As seen in the results, the random forest and gradient boosting models had a significantly better performance compared to the other models, which was reflected in the good balance between precision and recall for each of the risk levels. This corroborates the reputation of these models to capture complex data patterns. Furthermore, an ensemble model that combined the random forest and gradient boosting models, provided an even higher performance. This highlights the benefits that we can get from merging complementary models instead of using individual ones. Finally, I think this project shows how integrating machine learning into health-related problems can deliver strong predictive power even by studying a small but effective amount of predictors. This proves that

data-driven methods can benefit prenatal care programs by enhancing early risk detection, reducing complications, and possibly saving lives.

6. Future Work

I think for future work, a deeper exploratory data analysis could be performed by incorporating more significant features like past obstetric history, nutritional status, or even ultrasound measurements. Additionally, I could try to add feature engineering methods such as combining the systolic and diastolic blood pressure into median arterial pressure and study if this provides more accurate information. I also think I could try to enhance hyperparameter tuning by using more advanced optimization techniques like randomized search or Bayesian optimization. Finally, I could try more complex models like XGBoost, SVM, or neural networks to study if more sophisticated methods provide better insights into maternal health risk.