From Data to Delivery: Predicting Pregnancy Risk with a Boosted Decision Tree

McKenzie Skrastins and Leemor Waldman

December 2024

Abstract

Maternal health is a critical global priority, with an estimated 213 million pregnancies annually. High-risk pregnancies, defined as those with increased odds of complications, contribute significantly to maternal and neonatal mortality, particularly in low-income regions. In the United States alone, 65,000 high-risk pregnancies occur annually, with 80% of related deaths deemed preventable. This study aimed to create a statistical learning model that could classify the risk level of pregnancies occurring in Bangladesh. The data used contained 1,014 observations collected from maternity clinics in Bangladesh, encompassing key physiological features such as age, blood pressure, blood sugar, body temperature, heart rate, and risk level.

After pre-processing the dataset, a benchmark logistic regression model was created. The logistic regression achieved 65% accuracy, indicating relationships between features and risk levels. Due to the high bias, low variance nature of the dataset, a decision tree, specifically a boosted decision tree, was used to make the final predictions. A baseline decision tree model with optimized hyperparameters reached an accuracy of 82.3% and a boosted decision tree with tuned hyperparameters yielded a 10-fold cross validated accuracy of 85.81%. Furthermore, blood sugar level was found to be the most important predictor in classifying pregnancies

These results highlight the usefulness of statistical learning in identifying high-risk pregnancies, and emphasize the need for enhanced prenatal care, particularly in socioeconomically disadvantaged regions.

Introduction

With an estimated 213 million pregnancies occurring around the world each year, prioritizing maternal health is crucial in ensuring safe and successful pregnancies. The National Academies of Sciences et al. defines a 'high risk pregnancy' as a "*situation in which the pregnant woman, fetus, or both have an increased likelihood or odds of a pregnancy complication, adverse event, or poor outcomes occurring during or after the pregnancy or birth.*"[2]. Holness outlines that The Centers of Disease Control and Prevention estimate that in the United States, 65,000 women have high risk pregnancies, annually[3]. The Maternal Mortality Review Committees stated that 80% of pregnancy related deaths in the United States were preventable[4]. Furthermore, The World Health Organization states that in 2020, 95% of maternal deaths occurred in low and lower-middle income countries, specifically in Sub-Saharan African and Southern Asian countries[5], such as Bangladesh. 41.6% of pregnancies in Bangladesh were classified as high risk, and resulted in multiple health complications, says the National Institute of Health[6]. Bangladesh is one of the most populated countries in the world, and is currently undergoing major economic growth[7]. These sources outline the demand for the increase in the quality prenatal services available, especially to those who come from a lower socioeconomic background.

Ahmed et al conducted a series of surveys at hospitals and maternity clinics in Dhaka and Khulna, Bangladesh. Through the use of Radio-Frequency Identification sensors, data was collected and recorded on pregnant mothers. Ahmed et al assessed which factors lead to a high risk pregnancy[8]. The collected data consists of 1,014 observations, with 26.9% classified as high-risk pregnancies, 33.1% as mid-risk pregnancies, and 40.0% as low-risk pregnancies. Additionally, the dataset has 7 features, 'Age,' 'SystolicBP,' 'DiastolicBP,' 'BS,' 'BodyTemp,' 'HeartRate,' and 'Risk Level':

- Age: The number of years an individual has been alive.
- SystolicBP (Systolic Blood Pressure): The upper value of a blood pressure reading, measured in millimeters of mercury (mmHg). A healthy individual's Systolic Blood Pressure typically ranges between 90-120 mmHg.
- DiastolicBP (Diastolic Blood Pressure): The lower value of a blood pressure reading, measured in millimeters of mercury (mmHg). A healthy individual's Diastolic Blood Pressure typically ranges between 60-80 mmHg.
- BS (Blood Sugar): The amount of glucose in the blood stream, measured in molar concentration (mmol/L). A healthy individual's blood glucose levels are less than or equal to 7.8 mmol/L.
- BodyTemp (Body Temperature): The internal temperature of an individual, measured in Fahrenheit degrees. The typical body temperature for a healthy individual is 98.6 degrees Fahrenheit.

- HeartRate (Resting Heart Rate): The amount of times an individual's heart beats in one minute. The typical resting heart rate of a healthy adult is between 60 and 100 beats per minute.
- Risk Level: The predicted risk intensity level during pregnancy. The Risk Level classifies patients into three categories: 'low risk,' 'mid risk,' or 'high risk'.

If a thorough evaluation on the data can be performed, and a model that correctly classifies the observations as 'low risk,' 'mid risk,' or 'high risk' can be made, then alternative methods to encourage patients to seek prenatal care can be created.

Methods

The data were prepared by loading them into the programming environment as a *Pandas' DataFrame*. The proposed predictor, "RiskLevel", was removed from the dataframe and set equal to the variable y. The remaining dataframe was set equal to the variable X. Using scikit-learn's *train_test_split* function, the dataframe was divided into:

$$\begin{gathered} X_{\text{train}} = \{x_i \in X \mid i \in \text{train indices}\} \\ y_{\text{train}} = \{y_i \in y \mid i \in \text{train indices}\} \\ X_{\text{test}} = \{x_i \in X \mid i \in \text{test indices}\} \\ y_{\text{test}} = \{y_i \in y \mid i \in \text{test indices}\} \\ |X_{train}| = 0.8 * N \\ |y_{train}| = 0.8 * N \\ |X_{test}| = 0.2 * N \\ |y_{test}| = 0.2 * N \end{gathered}$$

where N is the total number of data points in X and y.

Therefore, 80% of the X and y data were put into a train set, and the remaining data (20%) into a test set. This process yielded four sets, $X_{train}, y_{train}, X_{test}$, and y_{test}, as stated above. The *random_state* argument was used to ensure that if the coding cell was run again, the split would remain the same. Finally, the class names in y were changed from, 'low risk,' 'mid risk,' and 'high risk' to 0, 1, and 2, respectively.

A benchmark logistic regression (LR) model was applied to the dataset to assess whether there was a relationship between the features and the predicted values. First, scikit-learn's *LogisticRegression* function was loaded in. Then, the LR model was fit to the X_{train} and y_{train} data. Finally, the X_{test} data was fed into the model and was used to make predictions, $\hat{y}$. To obtain an accuracy score, y_{test} and $\hat{y}$ were compared using scikit-learn's *accuracy_score* function. The accuracy score of the LR model was interpreted as follows: if

$$accuracy_{BenchmarkLR} \geq 0.57,$$

then, a relationship between X and y does exist, and it is possible to create a more complex and robust model to make predictions. The benchmark score of 0.57 was selected because it is slightly higher than guessing (0.50), but not so high that the model would be expected to perform well.

A preliminary analysis was conducted to determine whether the data would result in a high bias-low variance model or a low bias-high variance model. Principal Component Analysis (PCA) and Lasso Regression (L_1 Reg.), both dimensionality reduction techniques aimed at improving low bias-high variance models, were applied to the baseline logistic regression model. Ridge Regression (L_2 Reg.), a weight penalization method that also refines low bias-high variance models, was similarly applied. The analysis concluded that if these techniques improved the model's accuracy, the data would likely produce a low bias-high variance model. Conversely, if accuracy did not improve, the data would indicate a high bias-low variance model. Which model to proceed with was selected based on this classification.

After assessing several baseline classification methods and parameter tuning strategies, ensemble trees and boosting were the models selected. Examples of the baseline methods and the parameter tuning strategies explored include K-nearest-neighbors, support vector machines, random forests, and bagging.

A baseline tree was created using scikit-learn's *DecisionTreeClassifier* function. The model was loaded into the coding environment, and then fit to the X_{train} and y_{train} sets. The *random_state* argument was used to ensure that if the coding cell was run again, the model would remain the same. The X_{test} data was fed into the model and was used to make predictions, $\hat{y}$. To obtain an accuracy score, y_{test} and $\hat{y}$ were compared using scikit-learn's *accuracy_score* function.

Additionally, a tree depth and maximum nodes finder was made using scikit-learn's *GridSearchCV* function. 10-fold cross-validation splits were set, using scikit-learn's *KFold* function. *GridSearchCV* was loaded into the

coding environment and the DecisionTreeClassifier model was passed as an argument during the construction of the GridSearchCV model. The GridSearchCV model was fit to the X_{train} and y_{train} sets, and GridSearch's *best_params_* and *best_estimators_* methods were used to find the tree depth and maximum nodes. Scikit-learn's *cost_complexity_pruning_path* function was used to prune the tree and find the maximum nodes. The same steps used to find the tree depth were used to prune the tree; *cost_complexity_pruning_path* was loaded into the environment, the Tree model was passed as an argument into the construction of the pruning model, and the data were fit.

To construct the boosted tree model, the *GradientBoostingClassifier* function was used. *GradientBoostingClassifier* was introduced to the environment, fit to the X_{train} and y_{train} sets, and then made predictions on X_{test}. Next, a custom hyperparameter finder was used to identify the optimal learning rate and number of estimators. The following sets of hyperparameters were tested:

$$learn\_rates = [0.001, 0.01, 0.1, 1]$$
$$estimators = [900, 1000, 1250, 1500, 1750, 2000]$$

The boosting model was then recreated with the optimal learn rate and estimator size. The tree depth and maximum node, found from *best_params_* and *best_estimators_*, were applied to the Boosted models.

To visualize the performance of the model, a confusion matrix of the test data was created using scikit-learn's *confusion_matrix* function, and seaborn's *sns* class and *heatmap* method. To asses the validity of the model, 10-fold cross validation was conducted using scikit-learn's *KFold* function. The model was introduced to the environment with the arguments *shuffle* set to true and *random_state*. Additionally, along with an accuracy score, the precision, recall, and F1 scores of the model were calculated using scikit-learn's *precision_score*, *recall_score*, and *f1_score* functions. The Type I error and Type II error were also measured.

Results

Figure 1 illustrates the Logistic Regression model made. Note that for visualization purposes, one feature ('Blood Sugar') was selected, and the other features were left constant (averaged). See figure 1;

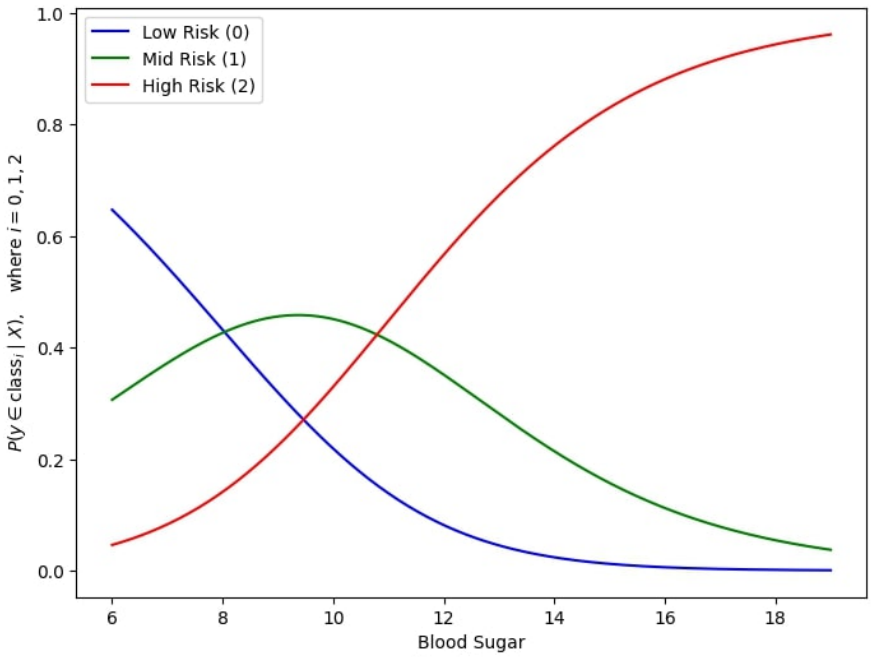

Figure 1: Plot of Logistic Regression Model, with $X = X_0 =$ Blood Sugar and where $X_1, ..., X_N$ are Averaged

The Benchmark Logistic Regression model yielded an accuracy score of

$$accuracy_{BenchmarkLogisticRegression} = 0.6502463054187192$$

The preliminary analysis, using PCA, L_1 Reg. and L_1 Reg., yielded the following accuracy scores;

$$accuracy_{BenchmarkLRWithPCA} = 0.5763546798029556$$
$$accuracy_{BenchmarkLRWithL1} = 0.625615763546798$$
$$accuracy_{BenchmarkLRWithL2} = 0.6502463054187192$$

Next, the Baseline Decision Tree was made. See figures 2 and 3;

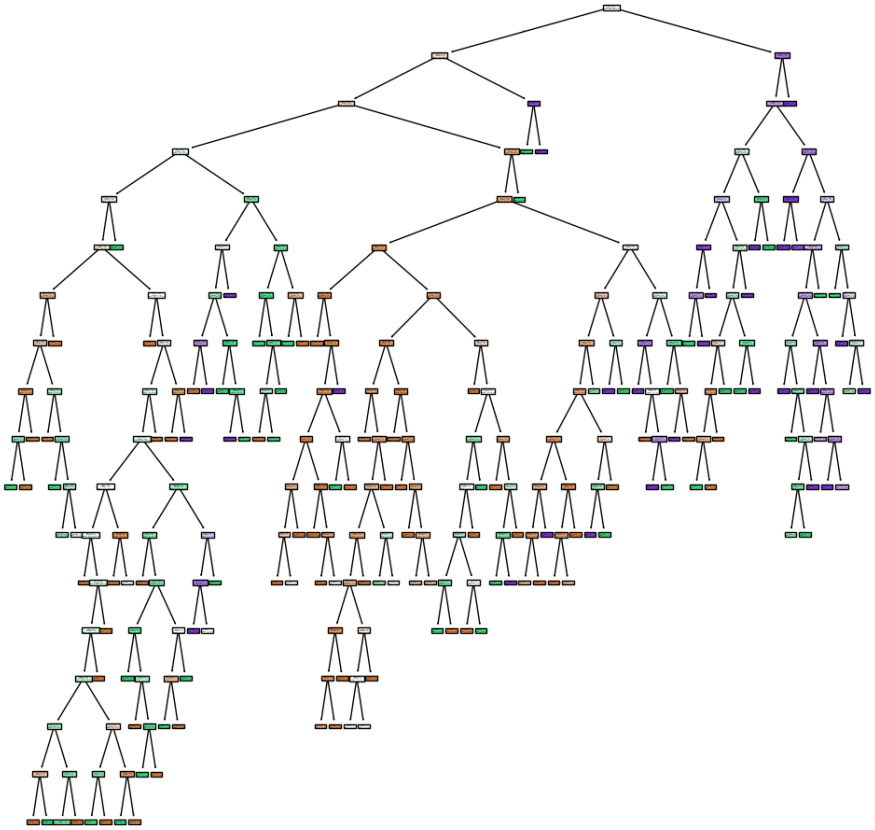

Figure 2: Baseline Ensemble Tree, with a Maximum Node count of 293 and a Maximum Depth of 17

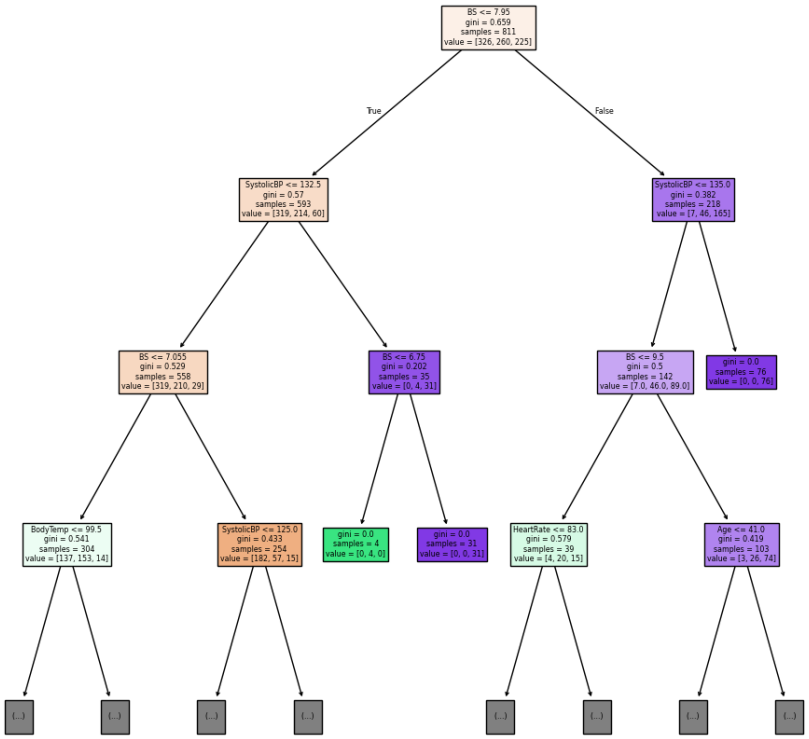

Figure 3: Baseline Ensemble Tree, with a Maximum Depth of 3

The Tree depth and maximum nodes finder assessed that;

$$Best\ Max\ Depth = 17$$
$$Max\ Nodes = 293$$

The accuracy of the Baseline Decision Tree, with these hyperparameters, was found to be:

$$accuracy_{BaselineDecisionTree} = 0.8226600985221675$$

The first Boosted model yielded an accuracy score of:

$$accuracy_{UntunedHyperparameterBoostedTree} = 0.8177339901477833$$

A custom hyperparamter finder was used to detect the best learning rate and estimator amount. The hyperparameters were found to be;

$$n\_estimators = 1500$$
$$learn\_rate = 0.01$$

Another Boosted model was created, taking into account the optimal hyperparameters. The accuracy of the second Boosted model was found to be;

$$accuracy_{TunedHyperparametersBoostedTree} = 0.8325123152709359$$

Additionally, the Confusion Matrix was made. See figure 4;

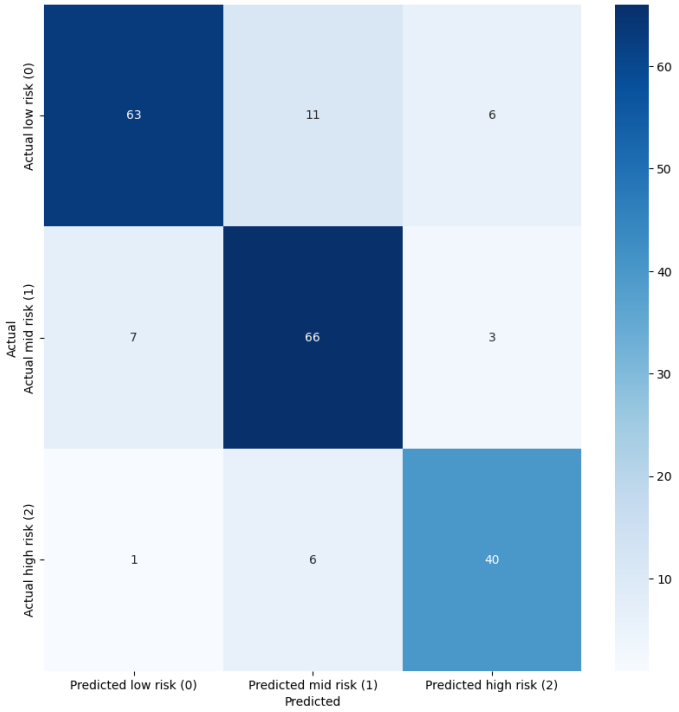

Figure 4: Confusion Matrix of Boosted Tree Model Predictions

After performing 10-fold cross validation, the following metrics were found:

$$10 - Fold\ Cross\ Validation\ Scores =$$
$$[0.81372549, 0.81372549, 0.83333333, 0.89215686, 0.84158416, 0.87128713, 0.91089109, 0.88118812, 0.83168317, 0.89108911]$$
$$Accuracy_{MeanOver10Folds} = 0.8580663948747815$$
$$Standard\ Deviation_{AccuracyOver10Folds} = 0.03356218800206466$$

Finally, precision, recall, F1, Type I and Type II error were calculated. See below;

$$precision_{TunedHyperparametersBoostedTree} = 0.8320681931309108$$
$$recall_{TunedHyperparametersBoostedTreeClass_0} = 0.7875000000000000$$
$$recall_{TunedHyperparametersBoostedTreeClass_1} = 0.8684210500000000$$
$$recall_{TunedHyperparametersBoostedTreeClass_2} = 0.8510638300000000$$
$$F1_{TunedHyperparametersBoostedTree} = 0.8326000000000000$$
$$Test\ I\ Error_{TunedHyperparametersBoostedTree} = 0.16748768472906406$$
$$Test\ II\ Error_{TunedHyperparametersBoostedTree} = 0.16793180686908915$$

To help with model interpretation, the level of importance of each predictor was found. See figure 5;

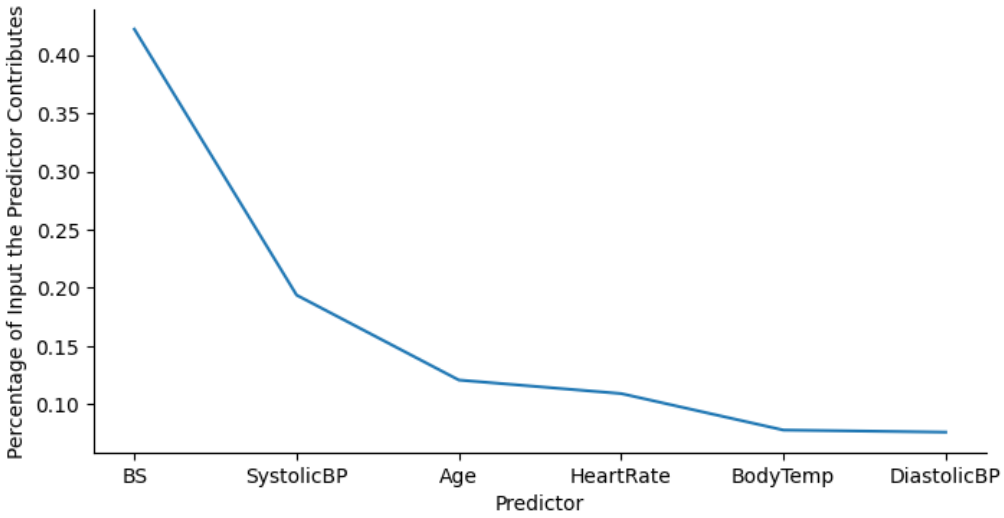

Figure 5: Importance of Predictors in the Boosted Tree Model

Discussion

This analysis aimed to develop a model capable of classifying pregnancies into three categories— 'low risk,' 'mid risk,' or 'high risk' —based on commonly observed factors during pregnancy. First, the Benchmark Logistic Regression (LR) model was made. The accuracy score of the LR model was objectively not optimal. However, the conclusion that the Model was not merely guessing could be drawn, as the obtained accuracy of 65% exceeded the threshold set of 57%. Therefore, a more robust and complex model could be created to make accurate predictions.

The preliminary analysis, using PCA, L_1 Reg. and L_2 Reg. produced accuracy scores that were either worse than or equivalent to the accuracy of the Benchmark Logistic Regression. Consequently, the data was projected to most likely result in a high bias-low variance model. This conclusion aligns with the characteristics of the dataset, which contains only six predictors. A small number of predictors reduces the likelihood of high variance but increases the risk of high bias. Therefore, it was concluded that a model capable of effectively managing and mitigating bias, such as a boosted decision tree, would be best.

Before creating the Boosted Decision Tree, the pruned Baseline Decision Tree was created, and the optimal node count and tree depth were determined. Despite the tree's size—293 nodes and a depth of 17—typically suggesting overfitting, the preliminary analysis proved that overfitting, or high variance, did not need to be a concern throughout this study. Figure 3 highlighted three leaf nodes with a perfectly pure gini index, indicating that all classifications at these nodes fell into a single class: low risk, mid risk, or high risk pregnancies. This purity demonstrated that the model successfully identified and leveraged patterns within the data to make accurate classifications.

Both the Untuned Boosted Tree and the Tuned Boosted Tree were made, and yielded the accuracy scores of 81.77% and 83.25%, respectively. Although the difference is small, it was deemed significant, as the additional 2% translates to accurately categorizing 20 more pregnancies.

10-Fold cross validation displayed accuracy scores ranging from 70% to 92%. Several conclusions could be drawn from this. First, the data could be noisy or contain lots of outliers in some folds. Second, some classes could be overrepresented in some folds, while some are underrepresented in others, leading to the overfitting or underfitting of the Model.

The Confusion Matrix (CM) for the test data was generated, revealing that the majority of misclassifications occurred when low-risk pregnancies were incorrectly classified as mid-risk pregnancies, accounting for approximately 32% of the misclassifications. While this outcome is not ideal, the real-world impact of such misclassifications would be less detrimental than misclassifying a high-risk pregnancy, which could have more serious consequences.

Additional metrics were analyzed to evaluate the model's validity. The recall scores for the different risk classes showed slight variation, with the low-risk class achieving a recall of 78%, while mid-risk and high-risk classes had recall scores of 85% and 86%, respectively. This discrepancy might suggest that the model may struggle to accurately identify low-risk pregnancies compared to mid- and high-risk ones. However, a more plausible explanation

is that because the model is trained on over 100 additional low-risk samples, the lower recall score reflects the variability within these extra samples. Additionally, both Test I and Test II errors were relatively high, around 16.7%. Although misclassifications are not ideal, the real-world impact of misclassifying a low-risk pregnancy is less critical than the misclassification of a mid or high-risk pregnancy. Further hyperparameter tuning and exposer to more mid and high risk samples are recommended to improve the model's performance.

As illustrated in Figure 4, 'Blood Sugar' consistently emerged as the most significant variable in predicting the risk level of pregnant women, contributing 42% of the input. Blood sugar levels are a key factor in diabetes, a condition frequently associated with high-risk pregnancies. The model's identification of 'Blood Sugar' as a critical predictor aligns with existing literature, reinforcing the biological validity of the Boosted Tree Model. While 'Systolic Blood Pressure' ranked as the second most important variable, its contribution was only 19%, representing a notable decrease compared to Blood Sugar. 'Age', the third most influential factor, is commonly cited as a significant determinant in classifying high-risk pregnancies. However, the Boosted Tree Model found that 'Age' contributed only 12% to the prediction, which is considerably lower than what is typically suggested in the literature. In her contribution to *Female Health Across the Lifespan*, Holness identifies four primary categories of high-risk pregnancy factors, with 'Age' occupying a distinct category. These findings suggest that health professionals should prioritize further investigation into the role of blood sugar levels in high-risk pregnancies, with an increased focus on diabetes when assessing pregnancy risk. The remaining predictors — 'HeartRate', 'BodyTemp', and 'DiastolicBP' — contributed 10% or less and were not further analyzed in this discussion.

Conclusion

In summary, this analysis demonstrated the development and evaluation of a Boosted Tree model for classifying pregnancy risk. Starting with a Benchmark Logistic Regression model, the data was confirmed to possess the potential for predictive accuracy, despite its initial limitations. A preliminary analysis revealed a high bias - low variance model, which was expected given that the dataset has minimal features. The transition to decision tree and boosted tree models highlighted the complexity of the dataset, with the tuned boosted tree achieving a cross validated accuracy of 85.81%. Cross-validation underscored the variability in the dataset, while a confusion matrix revealed meaningful insights into misclassifications and their practical implications. The study also identified "Blood Sugar" as the most significant predictor, aligning with existing literature on the role of diabetes in high-risk pregnancies. These findings suggest that further refinement of the model, such as continuing to fine tune hyperparameters, and placing an emphasis on managing blood sugar could enhance clinical decision-making and improve outcomes in maternal health.

A model like this could be transformed into an interactive online quiz, providing an accessible tool for pregnant individuals in Bangladesh to evaluate their pregnancy risk. By making such an evaluation available online, access to prenatal care could be significantly expanded, especially for those who might not otherwise seek it. Example questions might include: 'How old are you?', 'Do you have diabetes?', and 'On a scale from 1 to 10, what is your daily physical activity level?'.

Although the latter two questions do not directly align with the model's predictors, observations could be inferred from the provided responses. For instance, if a person has diabetes, their estimated 'Blood Sugar' observation might be ≥ 11.1 (mmol/L), consistent with the elevated glucose levels typical for individuals with diabetes. Similarly, if someone reports a physical activity level of ≥ 7, their estimated 'Heart Rate' could be ≤ 80 (bpm), reflecting the inverse relationship between activity level and resting heart rate.

Upon completing the quiz, individuals classified as high risk would be encouraged to seek prenatal care. The quiz could include links to local obstetricians and other pregnancy-related support services to guide users towards appropriate resources.

The overarching aim of the Boosted Model would be twofold: to motivate expecting parents in Bangladesh to pursue prenatal care, especially those who might not have otherwise, and to educate them about their health options and available support systems throughout pregnancy.

Acknowledgments

The authors would like to sincerely thank Dr. Vladislav Kargin for his guidance and feedback throughout the development of this project.

References

[1] Ahmed M., Kashem M.A., Rahman M., Khatun S. *(2020) Review and Analysis of Risk Factor of Maternal Health in Remote Area Using the Internet of Things* (IoT). In: Kasruddin Nasir A. et al. (eds) InECCE2019. Lecture Notes in Electrical Engineering, vol 632. Springer, Singapore.
$https://link.springer.com/chapter/10.1007/978-981-15-2317-5_30$

[2] National Academies of Sciences, Engineering, and Medicine, et al. Birth Settings in America: Outcomes, Quality, Access, and Choice. National Academies Press, 2020.

[3] Holness, Nola. "High-Risk Pregnancy." Female Health Across the Lifespan, Elsevier, Philedelphia, PA, 2018, pp. 242–242.

[4] "Four in 5 Pregnancy-Related Deaths in the U.S. Are Preventable." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 19 Sept. 2022, www.cdc.gov/media/releases/2022/p0919-pregnancy-related-deaths.html.

[5] The World Health Organization. "Maternal Mortality." World Health Organization, World Health Organization, 26 Apr. 2024, www.who.int/news-room/fact-sheets/detail/maternal-mortality.

[6] Abedin, Sumaiya, and Dharma Arunachalam. "Maternal Autonomy and High-Risk Pregnancy in Bangladesh: The Mediating Influences of Childbearing Practices and Antenatal Care." BMC Pregnancy and Childbirth, U.S. National Library of Medicine, 22 Sept. 2020,
$https://pmc.ncbi.nlm.nih.gov/articles/PMC7510296/pdf/12884_2020_Article_3260.pdf$

[7] Basu, Kaushik, et al. "Why Is Bangladesh Booming?" Brookings, Brookings, 9 Mar. 2022, www.brookings.edu/articles/why-is-bangladesh-booming/.

[8] IoT based Risk Level Prediction Model for Maternal Health Care in the Context of Bangladesh, STI-2020, [under publication in IEEE] $https://ieeexplore.ieee.org/document/9350320$