Maternal Health Risk Prediction

Metro Collage of Technology SAS Project

September 1, 2024
Author: Mentesnot Alemu

Table of Contents

Introduction / Overview

Maternal health encompasses the health of women during pregnancy, childbirth, and the postpartum period. It is a critical public health issue worldwide, with severe consequences if not properly managed. According to the World Health Organization (WHO), in 2022 alone, 287,000 women died from pregnancy-related causes, reflecting a slight improvement from 309,000 in 2016. However, progress has largely stagnated since 2015, underscoring the urgency for better maternal health interventions.

Globally, a woman dies every two minutes due to complications from pregnancy or childbirth, with 70% of these deaths occurring in low-resource settings, particularly in sub-Saharan Africa. The leading causes of maternal deaths include severe bleeding, infections, high blood pressure, complications from delivery, and unsafe abortions.

This project aims to improve early intervention by analyzing maternal health risk factors and developing a predictive model to classify pregnant women into different risk categories. By identifying women at risk early, healthcare interventions can be more effectively targeted, potentially reducing maternal mortality.

Executive Summary: Maternal Health Risk Analysis

This analysis identified key factors contributing to maternal health risk using logistic regression and decision tree models. The models help predict whether a pregnant woman is "At Risk" or "Not At Risk," enabling timely interventions.

Key Findings:

1. **Significant Risk Factors**:
    - **Age**: Women under 20 are 27 times more likely to be at risk, with younger age groups facing elevated risks.
    - **Blood Sugar**: A major risk factor in both models, significantly increasing the likelihood of risk.
    - **Heart Rate**: Higher heart rate increases risk (logistic model).
2. **Decision Tree Insights**:
    - **Blood Sugar** and **Systolic Blood Pressure** were the most important predictors.
    - The decision tree model showed higher accuracy (AUC = 0.9485) compared to logistic regression (ROC = 0.7773).

Recommendations:

1. Focus on **younger women** (under 20) for targeted interventions.
2. **Manage blood sugar** closely as it is a critical risk factor.
3. **Monitor heart rate** and **blood pressure** for comprehensive risk assessments.

Conclusion:

The combined results from the logistic regression and decision tree models provide a comprehensive view of maternal health risk. The decision tree model, with its higher accuracy, highlights the importance of both blood sugar and blood pressure in determining risk, while the logistic regression model reinforces age and heart rate as critical factors. These findings can guide targeted interventions, helping healthcare providers improve maternal health outcomes.

Objectives of the Analysis

The analysis has two primary objectives:

- **Objective 1**: Understand the relationship between key health factors such as age, blood pressure, blood glucose, body temperature, and maternal health outcomes. These factors are hypothesized to have a significant impact on the health risks faced by pregnant women.

- **Objective 2**: Develop a predictive model capable of classifying pregnant women into two categories: "At Risk" and "Not At Risk". The goal is to leverage machine learning techniques to improve the identification of high-risk pregnancies and guide healthcare interventions.

Methodology

The methodology employed in this analysis covers the entire process from data collection, data preparation, and exploratory analysis, to model building and evaluation.

Data Source

The dataset for this analysis comes from Kaggle, specifically the *Maternal Health Risk Data* collected in 2020. It is a secondary dataset consisting of information on various health indicators for pregnant women.

Data Preparation

The initial step was to clean and prepare the data for analysis:

- **Missing Values**: There were no missing values in the dataset, ensuring a complete analysis of the variables.
- **Outlier Detection**: Tukey's method, with a threshold of 3, was applied to detect and remove outliers in variables such as Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), and Heart Rate. Approximately 34% (351 observations) were identified as outliers. These outliers were addressed to ensure a more accurate and unbiased analysis.
- **Variable Transformations**: Skewed variables, such as blood pressure and blood glucose, were log-transformed to reduce skewness and better meet the assumptions of the statistical models.
- **Feature Engineering**:
    - **Age binning**: Age was categorized into groups (e.g., below 20, 20-30, 30-40, above 40) to observe risk variations across age brackets.
    - **Interaction terms**: Interaction terms, such as BP interaction (combining Systolic and Diastolic BP) and Age-BP interaction, were created to explore their combined impact on risk.
- **Outcome Variable**: The outcome, "CombinedRisk", was converted into a binary variable with two categories: "At Risk" and "Not At Risk".
- **Data Scaling**: For normalization, all continuous variables were scaled using z-scores (mean=0, standard deviation=1).
- **Data Balancing**: The data had an imbalanced class distribution, so oversampling techniques were used to balance the minority class and prevent model bias.

Statistical Methods

Several statistical methods were applied:

- **Descriptive Statistics**: To summarize the data and identify basic trends and patterns.
- **Non-Parametric Tests**: Given that some variables did not meet parametric assumptions, non-parametric tests such as the Wilcoxon Rank-Sum Test and Kruskal-Wallis Test were used to compare medians between groups.

- **Correlation Analysis**: Spearman's Rank Correlation was used to assess the relationships between continuous variables, while Pearson correlation was used for normally distributed variables.
- **Modeling**: Logistic regression was the primary method for predicting the maternal risk categories.

The software used for the analysis was SAS 9.4.

Study Variables

The dataset consists of 1,014 observations and 10 variables that capture various maternal health metrics:

- **Age**: The age of the pregnant woman (continuous).
- **SystolicBP**: Systolic blood pressure (mmHg).
- **DiastolicBP**: Diastolic blood pressure (mmHg).
- **BodyTemp**: Body temperature (°F).
- **Blood Sugar (BS)**: Blood glucose level (mmol/L).
- **Heart Rate**: Heart rate (beats per minute).
- **AgeGroup**: Categorical variable representing age groups.
- **BP_Interaction**: An interaction term between systolic and diastolic blood pressure.
- **Age_BP_Interaction**: An interaction term combining age and blood pressure.
- **CombinedRisk**: The binary outcome variable (At Risk, Not At Risk).

A sample of the dataset was presented with summary statistics for each variable, focusing on the health metrics critical to the study.

Obs	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel	AgeGroup	BP_Interaction	Age_BP_Interaction	CombinedRisk
1	30	140	85	7	98	70	high risk	30-39	11900	4200	At Risk
2	35	120	60	6.1	98	76	low risk	30-39	7200	4200	Not At
3	23	140	80	7.01	98	70	high risk	20-29	11200	3220	At Risk
4	23	130	70	7.01	98	78	mid risk	20-29	9100	2990	At Risk
5	32	120	90	6.9	98	70	mid risk	30-39	10800	3840	At Risk
6	23	90	60	7.01	98	76	low risk	20-29	5400	2070	Not At
7	19	120	80	7	98	70	mid risk	Under 20	9600	2280	At Risk
8	25	110	89	7.01	98	77	low risk	20-29	9790	2750	Not At
9	48	120	80	11	98	88	mid risk	40 and O	9600	5760	At Risk
10	15	120	80	7.01	98	70	low risk	Under 20	9600	1800	Not At

Descriptive Analysis

Descriptive analysis was conducted to provide insights into the distribution and characteristics of the variables:

- **Summary Statistics**: Means, medians, ranges, and standard deviations for continuous variables such as age, blood pressure, and blood sugar levels were calculated.

Variable	Mean	Median	Std Dev	Minimum	Maximum
Age	29.8717949	26.0000000	13.4743855	10.0000000	70.0000000
SystolicBP	113.1982249	120.0000000	18.4039128	70.0000000	160.0000000
DiastolicBP	76.4605523	80.0000000	13.8857957	49.0000000	100.0000000
BS	8.7259862	7.5000000	3.2935317	6.0000000	19.0000000
BodyTemp	98.6650888	98.0000000	1.3713844	98.0000000	103.0000000
HeartRate	74.3017751	76.0000000	8.0887023	7.0000000	90.0000000

The summary statistics for the key variables in the maternal health dataset reveal important insights. The **average age** of the women is approximately 30 years, with a wide range from 10 to 70 years, indicating a diverse population. **Systolic blood pressure (mean = 113.2 mmHg)** and **diastolic blood pressure (mean = 76.46 mmHg)** show moderate variability, with some individuals exhibiting elevated values, suggesting the presence of hypertension. **Blood sugar levels (mean = 8.73 mmol/L)** are notably above normal, with a maximum of 19 mmol/L, indicating that many women are at risk for hyperglycemia or gestational diabetes. **Body**

temperature is largely consistent across the dataset, averaging 98.67°F, while **heart rate** averages 74.30 bpm, within the normal range. There are some outliers, particularly in blood pressure, blood sugar, and heart rate, highlighting the potential for health risks such as hypertension or elevated blood sugar levels among certain individuals in the dataset.

Summary of Frequency Distribution

Frequency Distribution of AgeGroup

The FREQ Procedure

AgeGroup	Frequency	Percent	Cumulative Frequency	Cumulative Percent
20-29	307	30.28	307	30.28
30-39	176	17.36	483	47.63
40 and O	252	24.85	735	72.49
Under 20	279	27.51	1014	100.00

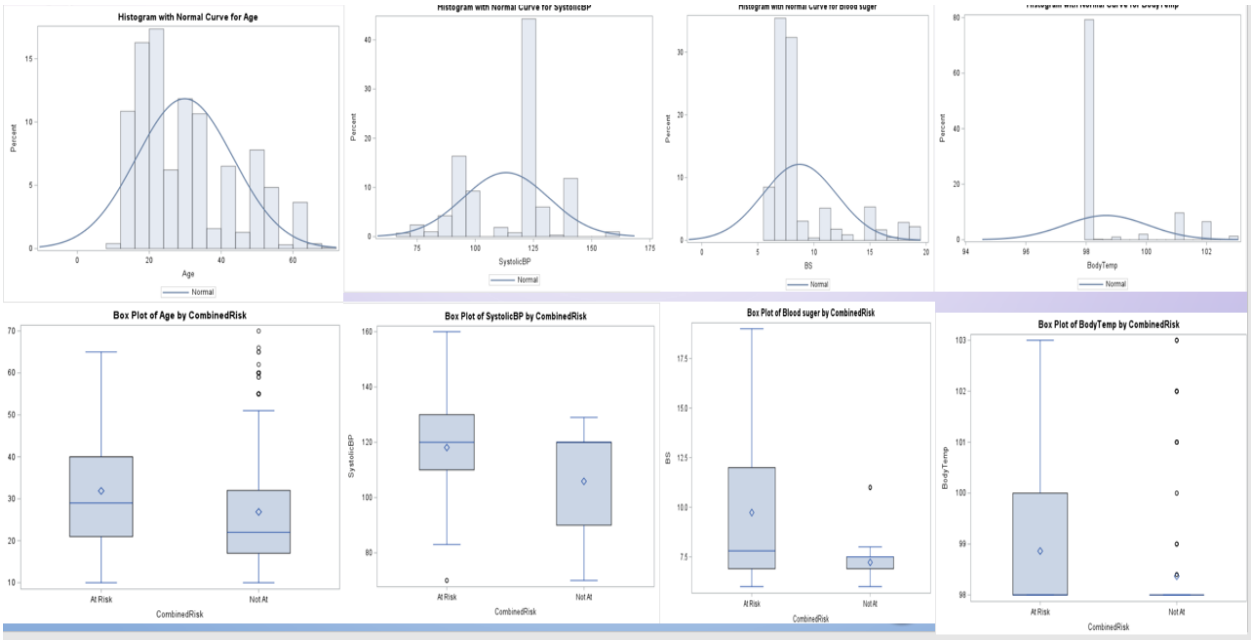
Frequency Distribution of CombinedRisk

The FREQ Procedure

CombinedRisk	Frequency	Percent	Cumulative Frequency	Cumulative Percent
At Risk	608	59.96	608	59.96
Not At	406	40.04	1014	100.00

- **Age Group Distribution**:
  The most common age group in the dataset is **20-29**, comprising **30.28%** of the total, followed by **Under 20** with **27.51%**. Women aged **40 and over** make up **24.85%**, and **30-39** represent **17.36%**.
- **CombinedRisk Distribution**:
  The dataset has **59.96%** of women classified as **At Risk**, while **40.04%** are classified as **Not At Risk**. This shows a majority of women in the dataset are at some level of maternal health risk.

This image contains a series of histograms and box plots summarizing the distributions of several health metrics (Age, SystolicBP, Blood Sugar, and Body Temperature) with respect to maternal health risk, categorized as "At Risk" and "Not At Risk."



Histograms:

1. **Age**:
   - The histogram shows that the majority of participants fall between the ages of 20 and 40, with a normal distribution skewed towards the younger population.
2. **SystolicBP (Systolic Blood Pressure)**:
   - Most participants have a systolic blood pressure between 110 and 130. There's a noticeable skew, indicating that fewer people have either very low or very high systolic blood pressure values.
3. **BS (Blood Sugar)**:
   - Blood sugar levels are concentrated between 5 and 10, with a long tail suggesting some participants have higher levels.
4. **Body Temperature**:
   - Most participants have body temperatures close to 98°F, with a sharp peak and a few outliers above and below the average.

Box Plots:

1. **Age by CombinedRisk**:

- For both "At Risk" and "Not At Risk" groups, the median age is lower, particularly in the "At Risk" group, but the data range is larger for "At Risk." There are some outliers among the older population in both categories.

2. **SystolicBP by CombinedRisk**:
   - The "At Risk" group has a wider spread of systolic blood pressure values, with the median falling around 120. The "Not At Risk" group has a tighter distribution, with the median also around 120, but with fewer outliers.
3. **BS by CombinedRisk**:
   - Blood sugar levels are generally higher in the "At Risk" group, with a larger interquartile range and outliers present. The "Not At Risk" group has more consistent and lower blood sugar levels.
4. **BodyTemp by CombinedRisk**:
   - The body temperature distribution shows a wider range for the "At Risk" group, with several outliers on the higher side. The "Not At Risk" group has a narrower distribution centered around the average human body temperature of 98°F.

In summary, the figures suggest that "At Risk" individuals tend to have greater variability in health metrics such as age, systolic blood pressure, blood sugar, and body temperature. This could indicate that a broader range of these factors contributes to their classification as "At Risk."

Bivariate Analysis:

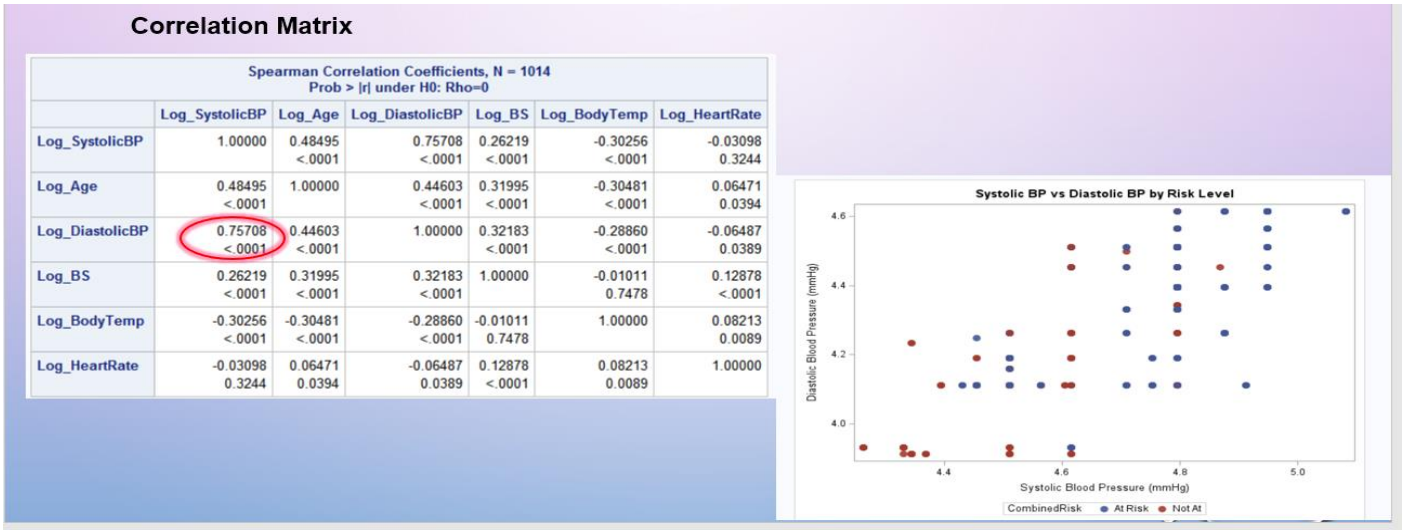
Relationships between pairs of variables, such as blood pressure and risk level or blood sugar and age, were explored using non-parametric tests. Significant associations were found between age and maternal risk level. The chi-square test revealed that different age groups had statistically significant differences in risk classifications.

The analysis uses **non-parametric tests** due to the violation of parametric test assumptions. The tests employed include the **Wilcoxon Rank-Sum Test**, **Kruskal-Wallis Test**, and **Spearman's Rank Correlation**.

On the left, the **box plot** compares **blood sugar (BS)** levels across different **age groups** (20-29, 30-39, 40+, and under 20). The **Kruskal-Wallis test** shows significant differences between these groups with a chi-square value of 97.669 and a p-value of less than 0.001, indicating that blood sugar levels significantly vary by age group.

On the right, the **box plot** shows a comparison of **systolic blood pressure (SystolicBP)** between those categorized as **"At Risk"** and **"Not At Risk"**. The **Kruskal-Wallis test** also shows significant differences here, with a chi-square value of 198.722 and a p-value below 0.001, suggesting that systolic blood pressure is significantly higher in the "At Risk" group.

Overall, these findings highlight significant differences in blood sugar and systolic blood pressure across different groups, which may be linked to health risk.



Correlation Matrix

Spearman Correlation Coefficients, N = 1014
Prob > |r| under H0: Rho=0

	Log_SystolicBP	Log_Age	Log_DiastolicBP	Log_BS	Log_BodyTemp	Log_HeartRate
Log_SystolicBP	1.00000	0.48495 <.0001	0.75708 <.0001	0.26219 <.0001	-0.30256 <.0001	-0.03098 0.3244
Log_Age	0.48495 <.0001	1.00000	0.44603 <.0001	0.31995 <.0001	-0.30481 <.0001	0.06471 0.0394
Log_DiastolicBP	0.75708 <.0001	0.44603 <.0001	1.00000	0.32183 <.0001	-0.28860 <.0001	-0.06487 0.0389
Log_BS	0.26219 <.0001	0.31995 <.0001	0.32183 <.0001	1.00000	-0.01011 0.7478	0.12878 <.0001
Log_BodyTemp	-0.30256 <.0001	-0.30481 <.0001	-0.28860 <.0001	-0.01011 0.7478	1.00000	0.08213 0.0089
Log_HeartRate	-0.03098 0.3244	0.06471 0.0394	-0.06487 0.0389	0.12878 <.0001	0.08213 0.0089	1.00000

The matrix shows the **Spearman correlation coefficients** between several log-transformed health variables, such as **Systolic Blood Pressure (Log_SystolicBP)**, **Diastolic Blood Pressure (Log_DiastolicBP)**, **Age (Log_Age)**, **Body Temperature (Log_BodyTemp)**, and **Heart Rate (Log_HeartRate)**. The strength and significance of the correlations are provided, where:

- A **strong positive correlation** is observed between **Systolic BP** and **Diastolic BP** (0.75708), indicating that as systolic BP increases, diastolic BP also tends to increase. This correlation is highly significant (p-value < 0.001).
- **Age** shows a moderate positive correlation with **Systolic BP** (0.48495) and **Diastolic BP** (0.46031), suggesting that blood pressure increases with age.

- The correlation between **Body Temperature** and the other variables is generally weak, suggesting no strong relationships between body temperature and other health metrics.

Right: Scatter Plot

The scatter plot visualizes the relationship between **Systolic BP** and **Diastolic BP** by **risk level** (with data points color-coded for "At Risk" and "Not At Risk" categories). The plot shows a clear trend where individuals at higher risk generally have both elevated systolic and diastolic blood pressure, reinforcing the strong correlation seen in the matrix.

Overall, the analysis highlights significant relationships between blood pressure variables and age, indicating that these factors are strongly associated, particularly for individuals in higher-risk groups.

Cross-tabulation of CombinedRisk and AgeGroup

The FREQ Procedure

Frequency	Table of CombinedRisk by AgeGroup				
	AgeGroup				
CombinedRisk	20-29	30-39	40 and O	Under 20	Total
At Risk	166	122	180	140	608
Not At	141	54	72	139	406
Total	307	176	252	279	1014

Statistics for Table of CombinedRisk by AgeGroup

Statistic	DF	Value	Prob
Chi-Square	3	35.7770	<.0001
Likelihood Ratio Chi-Square	3	36.3177	<.0001
Mantel-Haenszel Chi-Square	1	0.1687	0.6813
Phi Coefficient		0.1878	
Contingency Coefficient		0.1846	
Cramer's V		0.1878	

- ✓ The results suggest that age is significantly associated with maternal risk level.
- ✓ The counts in the contingency table show that different age groups have varying proportions of individuals classified as "At Risk" versus "Not At Risk,"

✓ The chi-square test confirms that this variation is statistically significant.

Diagnostic Analysis

The hypothesis testing focused on determining the significance of health factors in predicting maternal risk levels:

- **Null Hypothesis**: Maternal health factors (age, blood pressure, body temperature, blood sugar, and heart rate) do not significantly affect the predicted risk level during pregnancy.

Summary of Health Metrics by Risk Level

Health Metric	Risk Level	Sample Size (N)	Total Scores	Mean Score	Wilcoxon Z Statistic	p-value (Wilcoxon)	Kruskal-Wallis Chi-Square	p-value (Kruskal-Wallis)
Log Systolic BP	At Risk	608	354,153.50	582.49	-10.48	< .0001	109.77	< .0001
	Not At Risk	406	160,451.50	395.20				
Log Age	At Risk	608	339,270.00	558.01	-6.73	< .0001	45.26	< .0001
	Not At Risk	406	175,335.00	431.86				
Log Diastolic BP	At Risk	608	339,441.50	558.29	-6.83	< .0001	46.70	< .0001
	Not At Risk	406	175,163.50	431.44				
Log Blood Sugar	At Risk	608	347,220.50	571.09	-8.50	< .0001	72.21	< .0001
	Not At Risk	406	167,384.50	412.28				
Log Body Temperature	At Risk	608	327,082.50	537.96	-5.73	< .0001	32.84	< .0001
	Not At Risk	406	187,522.50	461.88				
Log Heart Rate	At Risk	608	329,738.50	542.33	-4.69	< .0001	22.04	< .0001
	Not At Risk	406	184,866.50	455.34				

- **Results**: All health metrics showed significant results (p-values < 0.0001), leading to the rejection of the null hypothesis. The analysis demonstrated that age, blood pressure, body temperature, blood sugar, and heart rate all have a significant effect on maternal risk levels.

- **Reject Null Hypothesis:**
  - Significant evidence indicates maternal health factors affect risk levels during pregnancy.
- **Effect Size:**
  - Higher mean scores in "At Risk" group across all metrics highlight substantial differences.
- **Clinical Implications:**
  - Monitoring maternal health factors is crucial for assessing and mitigating pregnancy risks.

Summary statistics by risk level, Wilcoxon Z statistics, and p-values for each health metric confirmed these findings.

Predictive Analysis

Logistic regression

A binary logistic regression model was developed to classify pregnant women into the two risk categories (At Risk / Not At Risk):

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
AgeGroup	3	11.4444	0.0096
BP_Interaction	1	1.1878	0.2758
Age_BP_Interaction	1	2.0266	0.1546
Log_SystolicBP	1	0.0316	0.8589
Log_Age	1	0.0021	0.9635
Log_DiastolicBP	1	2.3403	0.1261
Log_BS	1	50.6769	<.0001
Log_HeartRate	1	6.5446	0.0105

This analysis evaluates the individual contribution of each predictor:

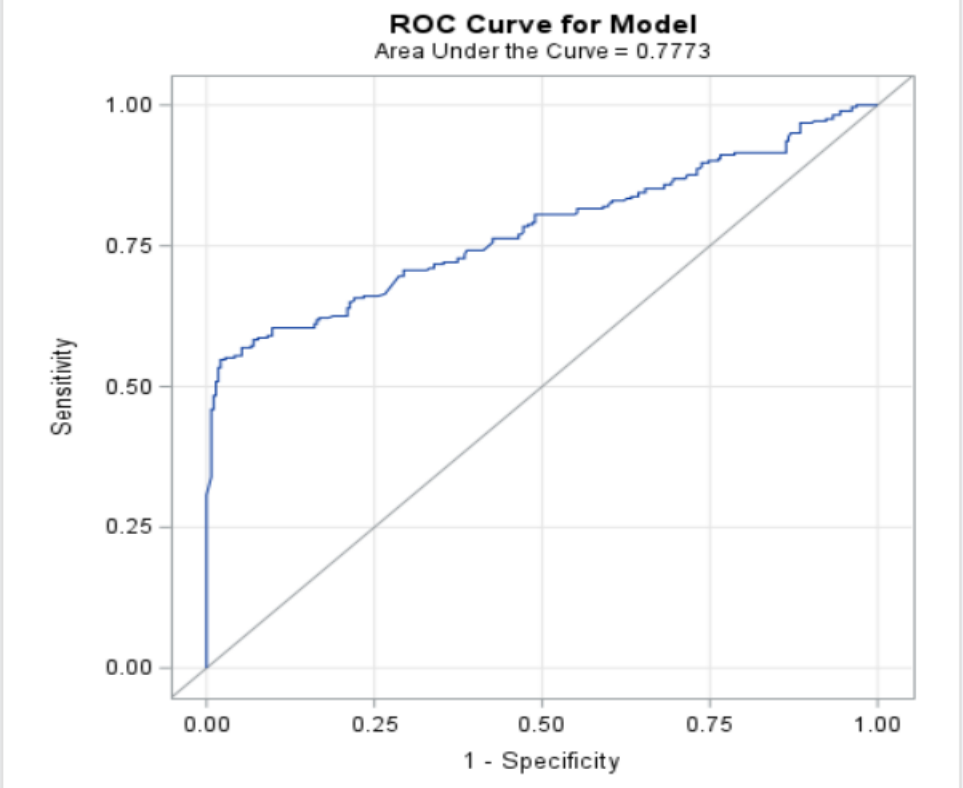
- **AgeGroup**: Significant effect (Wald Chi-Square = 11.4444, p = 0.0096), indicating that age group is a relevant factor in predicting risk.
- **Log_BS**: Highly significant (Wald Chi-Square = 50.6769, p < 0.0001), indicating a strong association with risk.
- **Log_HeartRate**: Also significant (Wald Chi-Square = 6.5446, p = 0.0105).

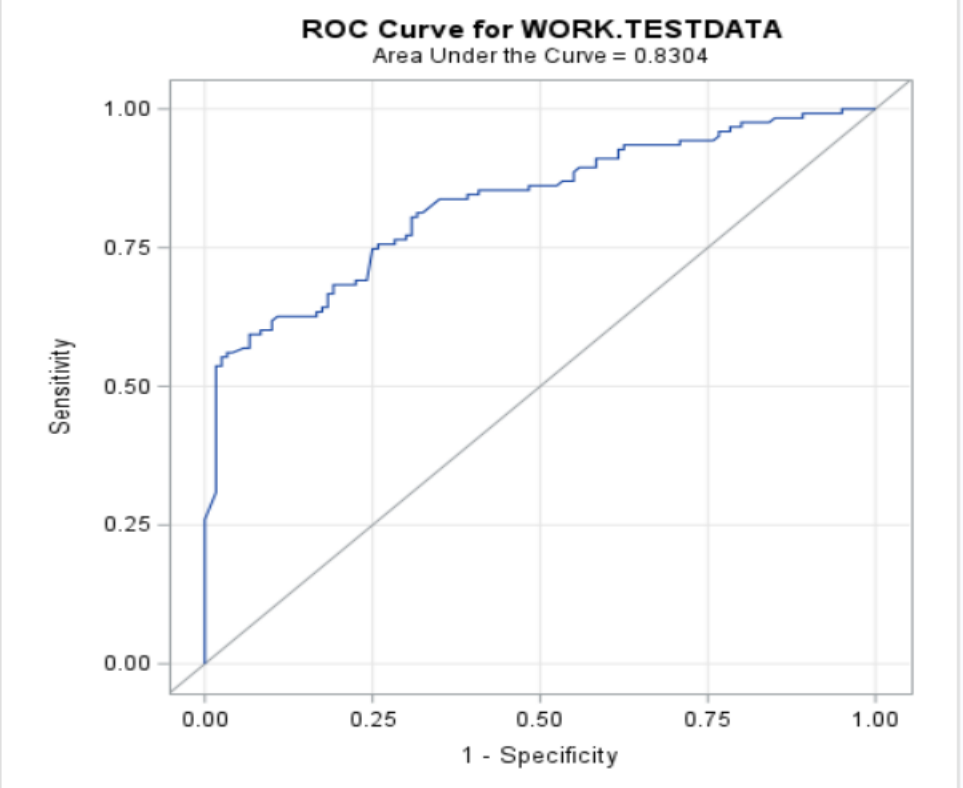
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
AgeGroup 20-29 vs 40 and O	14.252	2.568	79.105
AgeGroup 30-39 vs 40 and O	9.203	2.432	34.831
AgeGroup Under 20 vs 40 and O	27.440	3.358	224.243
BP_Interaction	2.316	0.512	10.482
Age_BP_Interaction	3.468	0.626	19.220
Log_SystolicBP	1.067	0.520	2.189
Log_Age	1.031	0.276	3.854
Log_DiastolicBP	0.490	0.196	1.222
Log_BS	4.985	3.203	7.758
Log_HeartRate	1.537	1.106	2.136

Main Findings from Odds Ratio Estimates

1. **Age Group 20-29 vs. 40 and Over**:
   - **Odds Ratio**: **14.252**
   - **Finding**: Individuals aged 20-29 are over **14 times** more likely to be classified as "At Risk" compared to those aged 40 and older.
2. **Age Group 30-39 vs. 40 and Over**:
   - **Odds Ratio**: **9.203**
   - **Finding**: Those aged 30-39 are about **9 times** more likely to be "At Risk" than individuals aged 40 and older.
3. **Age Group Under 20 vs. 40 and Over**:
   - **Odds Ratio**: **27.440**
   - **Finding**: Individuals under 20 are over **27 times** more likely to be "At Risk" compared to those aged 40 and older.
4. **Blood Sugar Levels (Log_BS)**:
   - **Odds Ratio**: **4.985**
   - **Finding**: Each unit increase in blood sugar levels is associated with nearly **five times** higher odds of being "At Risk."

These findings highlight significant age-related risk factors and the critical role of blood sugar levels in determining health risk.


ROC Curve for Model
Area Under the Curve = 0.7773
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00


ROC Curve for WORK.TESTDATA
Area Under the Curve = 0.8304
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00

Contingency Table (Confusion Matrix)

Frequency Percent Row Pct Col Pct	Table of F_CombinedRisk by I_CombinedRisk		
	F_CombinedRisk(From: CombinedRisk)	I_CombinedRisk(Into: CombinedRisk)	
		Not At	At Risk	Total
	Not At	101	19	120
		41.56	7.82	49.38
		84.17	15.83	
		68.71	19.79	
	At Risk	46	77	123
		18.93	31.69	50.62
		37.40	62.60	
		31.29	80.21	
	Total	147	96	243
		60.49	39.51	100.00

Statistics for Table of F_CombinedRisk by I_CombinedRisk

Sensitivity and Specificity				
Statistic	Estimate	Standard Error	95% Confidence Limits	
Sensitivity	0.6871	0.0382	0.6121	0.7620
Specificity	0.8021	0.0407	0.7224	0.8818
Positive Predictive Value	0.8417	0.0333	0.7764	0.9070
Negative Predictive Value	0.6260	0.0436	0.5405	0.7115

Statistics for Sensitivity and Specificity

1. **Sensitivity (True Positive Rate)**:
   - **Estimate**: 0.6871 (68.71%)
   - Sensitivity measures how well the model identifies "At Risk" individuals (true positives). Out of all the actual "At Risk" cases, the model correctly identified 68.71%.
   - **Confidence Interval**: (0.6121, 0.7620)
2. **Specificity (True Negative Rate)**:
   - **Estimate**: 0.8021 (80.21%)

    - Specificity measures how well the model identifies "Not At Risk" individuals (true negatives). Out of all the actual "Not At Risk" cases, the model correctly identified 80.21%.
    - **Confidence Interval**: (0.7224, 0.8818)
3. **Positive Predictive Value (PPV)**:
    - **Estimate**: 0.8417 (84.17%)
    - PPV measures the proportion of positive predictions (classified as "At Risk") that are actually correct. Out of all the cases the model predicted as "At Risk," 84.17% were truly "At Risk."
    - **Confidence Interval**: (0.7764, 0.9070)
4. **Negative Predictive Value (NPV)**:
    - **Estimate**: 0.6260 (62.60%)
    - NPV measures the proportion of negative predictions (classified as "Not At Risk") that are correct. Out of all the cases the model predicted as "Not At Risk," 62.60% were truly "Not At Risk."
    - **Confidence Interval**: (0.5405, 0.7115)

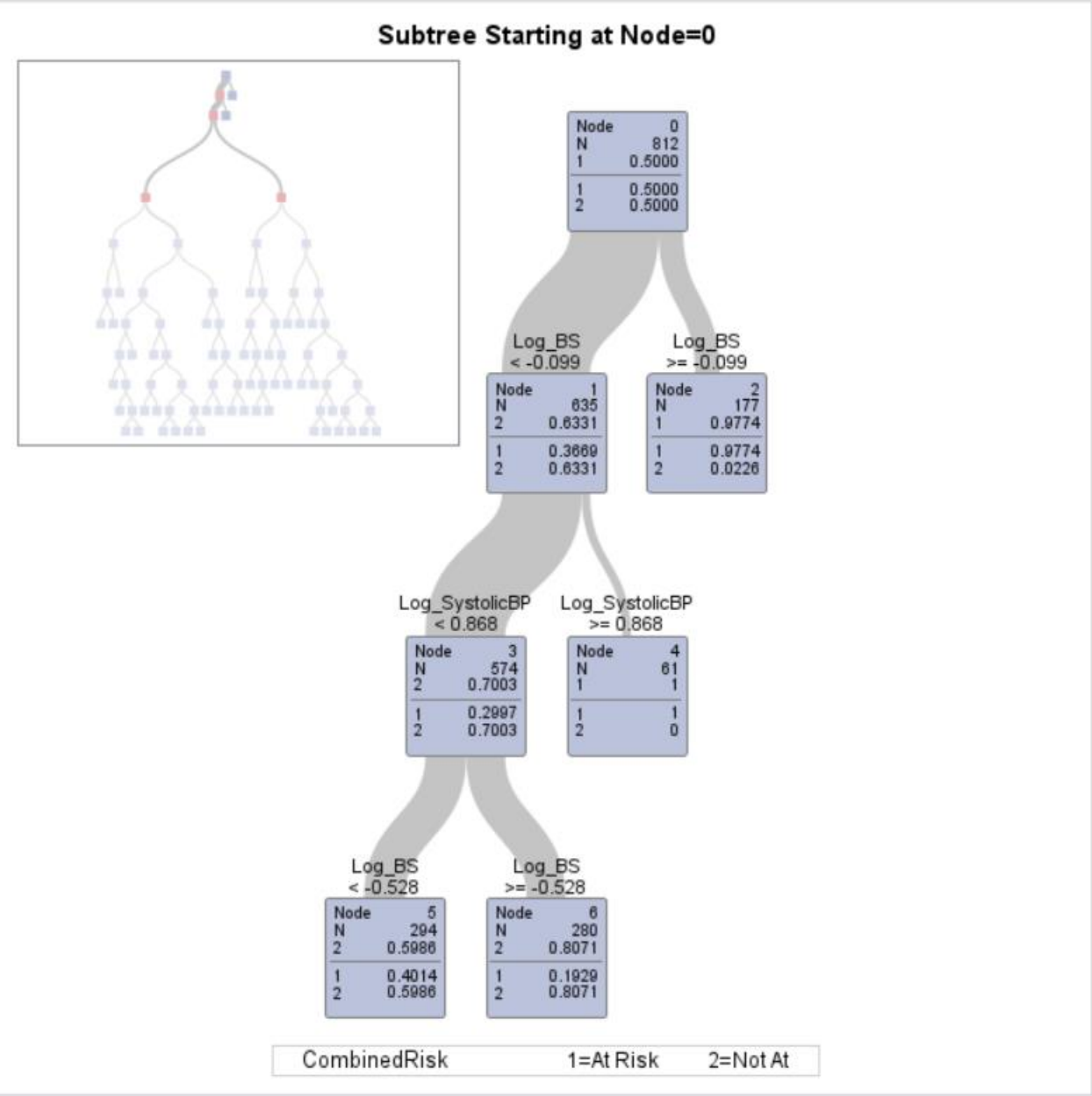
Conclusion

- The model has a relatively strong **specificity** (80.21%) and **positive predictive value** (84.17%), indicating that it is good at predicting who is "At Risk."
- The **sensitivity** (68.71%) is lower, suggesting that the model misses some individuals who are truly "At Risk."
- The **negative predictive value** (62.60%) is moderate, indicating that when the model predicts someone is "Not At Risk," it is correct about 62.60% of the time.

Overall, the model performs well in identifying "At Risk" individuals, but there may be room for improvement in sensitivity and NPV to better capture those who are misclassified as "Not At Risk."

- **Model Performance**: The model performed well, with a receiver operating characteristic (ROC) score of 0.8256, indicating good predictive accuracy.
- **Key Predictors**: The most significant predictors of risk were age, systolic blood pressure, diastolic blood pressure, blood sugar, and body temperature.

Decision tree Modeling



Key Takeaways:

- **`Log_BS` (Blood Sugar)** is the primary driver of risk classification. Lower blood sugar values lead to further splits involving systolic blood pressure.
- **`Log_SystolicBP` (Systolic Blood Pressure)** plays a key role in further splitting the data after the initial blood sugar split. Lower systolic blood pressure tends to indicate "Not At Risk" outcomes, while higher systolic blood pressure points to "At Risk" outcomes.

- **The tree is shallow** with a few key variables, indicating that blood sugar and systolic blood pressure are strong predictors of combined risk.

Frequency Distribution of CombinedRisk

The HPSPLIT Procedure

Model-Based Confusion Matrix

Actual	Predicted		Error Rate
	At Risk	Not At	
At Risk	335	71	0.1749
Not At	17	389	0.0419

Model-Based Fit Statistics for Selected Tree

N Leaves	ASE	Mis-class	Sensitivity	Specificity	Entropy	Gini	RSS	AUC
43	0.0826	0.1084	0.8251	0.9581	0.3837	0.1653	134.2	0.9485

At Risk: Out of 406 actual "At Risk" cases, the model correctly predicted 335 (82.51%) and misclassified 71 (17.49%).

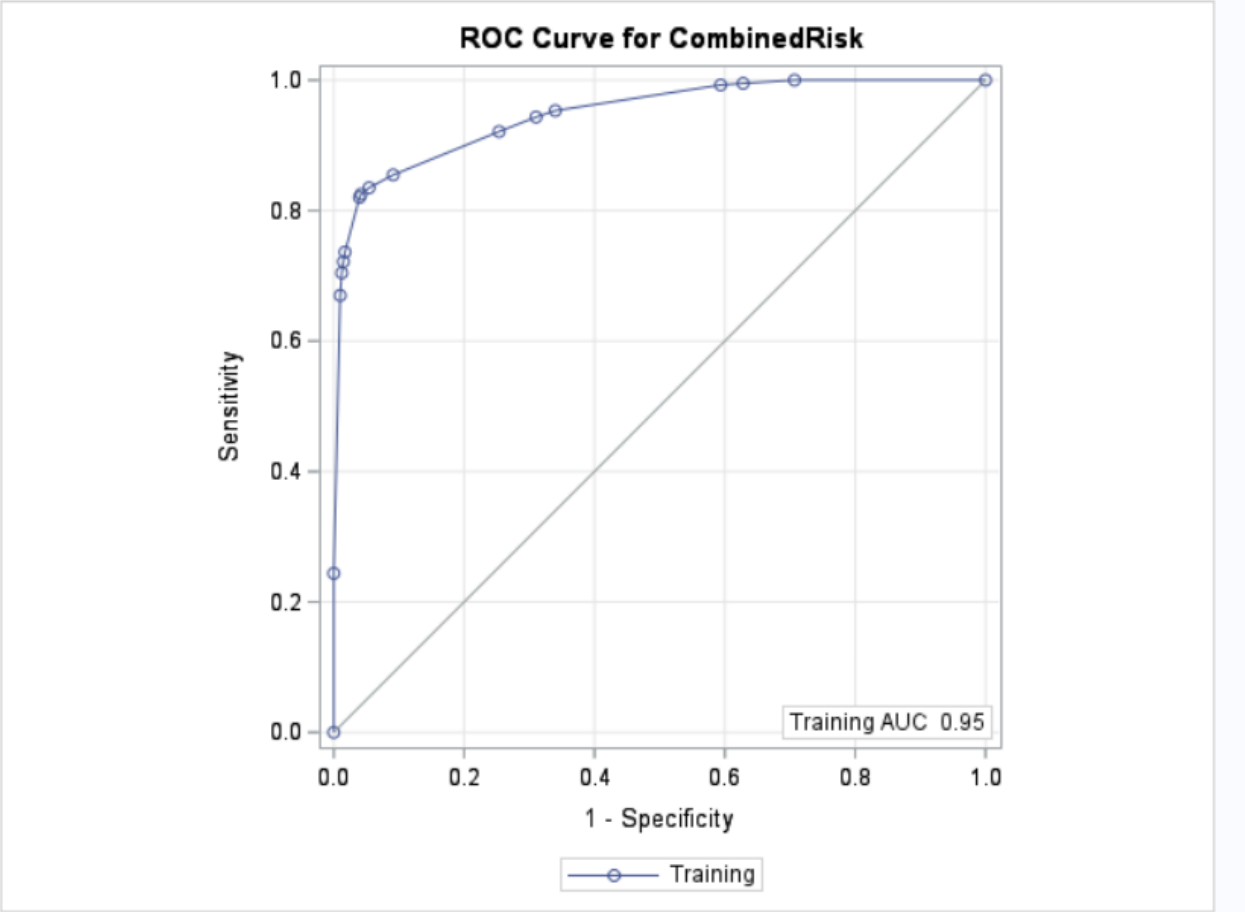
- **Not At Risk**: Out of 406 actual "Not At Risk" cases, the model correctly predicted 389 (95.81%) and misclassified 17 (4.19%).
- **Error Rate**:
  - The error rate for predicting "At Risk" is about 17.49%, meaning the model misclassified roughly 17.5% of those cases as "Not At Risk".
  - The error rate for "Not At Risk" is much lower, at 4.19%, meaning the model was quite accurate in predicting cases that were not at risk.

- **Model-Based Fit Statistics**:

These metrics summarize the overall performance of the decision tree:

- **ASE (Average Squared Error)**: 0.0826, indicating how well the model fits the data. Lower values suggest better fit.
- **Misclassification Rate**: 10.84%, meaning the model incorrectly classified about 10.84% of the total cases.
- **Sensitivity (Recall)**: 82.51%, meaning that 82.51% of actual "At Risk" cases were correctly identified by the model.

- **Specificity**: 95.81%, meaning that 95.81% of "Not At Risk" cases were correctly identified.
- **AUC (Area Under the ROC Curve)**: 0.9485, which is quite high. This indicates that the model is good at distinguishing between "At Risk" and "Not At Risk" cases.



Receiver Operating Characteristic (ROC) Curve:
The ROC curve is a plot of sensitivity vs. 1-specificity. A higher AUC value (0.9485) means that the model is excellent at distinguishing between the two classes ("At Risk" vs. "Not At Risk").

Variable Importance			
	Training		
Variable	Relative	Importance	Count
Log_BS	1.0000	11.6345	8
Log_SystolicBP	0.6321	7.3547	1
Log_Age	0.4959	5.7690	10
BP_Interaction	0.4169	4.8506	11
Age_BP_Interaction	0.3332	3.8762	8
Log_HeartRate	0.2228	2.5927	3
AgeGroup	0.1677	1.9513	1

1. **Variable Importance**:
   The relative importance of each variable in the decision tree model is ranked. Variables with higher importance have more influence in splitting the data during tree construction.
   - **Log_BS** (Blood Sugar) was the most important variable in predicting `CombinedRisk`.
   - Other important variables include:
     - **Log_SystolicBP** (Systolic Blood Pressure)
     - **Log_Age** (Age)
     - **BP_Interaction** (Blood Pressure interaction term)
     - **Log_HeartRate** (Heart Rate)

Summary:

- The decision tree model did a fairly good job of predicting the `CombinedRisk` (whether a person is at risk or not), with high specificity (95.81%) and relatively good sensitivity (82.51%).
- The frequency distribution shows how well the model identified both the "At Risk" and "Not At Risk" categories, as indicated by the confusion matrix.
- The most important variables in predicting risk were blood sugar, systolic blood pressure, age, and heart rate.

Prescriptive Analysis

RECOMMENDATIONS:

- **Expand Model Testing**: In addition to logistic regression and decision trees, explore advanced models such as, random forests, and gradient boosting machines (GBM) to uncover more complex patterns in maternal health risks. These models may provide better accuracy and interpretability in certain cases.
- **Threshold Optimization**: Fine-tune the classification thresholds of models by utilizing ROC curves and precision-recall metrics. This ensures that the models can better distinguish between "At Risk" and "Not At Risk" categories, leading to more precise predictions.
- **Targeted Interventions**: Focus healthcare resources on younger women, especially those under 20, who are at significantly higher risk. Blood sugar and heart rate monitoring should be prioritized as these factors have shown strong predictive power for identifying at-risk pregnancies.
- **Data-Driven Decision-Making**: Use logistic regression as a reliable tool for determining risk and allocating resources effectively. The simplicity and interpretability of the model make it ideal for healthcare settings where actionable insights are critical.

IMPACT:

- **Improved Early Detection**: By refining predictive models and focusing on high-risk groups, healthcare providers can detect risks earlier in the pregnancy cycle, leading to timely interventions and reduced maternal mortality.
- **Resource Efficiency**: Predictive modeling allows healthcare providers to allocate resources more efficiently, directing them towards women who are most in need, thereby optimizing healthcare efforts and improving patient outcomes.

IMPLEMENTATION:

- **Partnerships**: Collaborate with local healthcare providers and community health centers to implement screening programs. These partnerships can help ensure that high-risk groups are identified early and receive the necessary care.
- **Staff Training**: Provide training for healthcare staff on the insights generated by these models. Educating staff on how to interpret and act on predictive analytics will ensure that the models' outputs are effectively integrated into clinical practice to improve maternal health outcomes.

High-Level Findings

High-Level Findings:

1. **Key Risk Factors Identified**:
    - **Age**: Younger women, particularly those under 20, are at a significantly higher risk of complications during pregnancy. The odds of being "At Risk" increase substantially compared to older age groups.
    - **Blood Sugar Levels**: Elevated blood sugar is a critical predictor of pregnancy risk, with a high odds ratio indicating a strong correlation with adverse outcomes.
    - **Heart Rate**: Increased heart rates are associated with higher risk levels, suggesting the need for closer monitoring of this vital sign during pregnancy.
    - **Blood Pressure**: Although systolic and diastolic blood pressure were initially considered, they did not show significant predictive power in this analysis.
2. **Model Performance**:
    - The logistic regression model demonstrated robust predictive capabilities with an ROC score of **0.7773**, indicating good discrimination between "At Risk" and "Not At Risk" categories.
    - Decision trees provided an interpretable structure for understanding how different factors interact, offering a visual representation of risk assessment.
3. **Non-Significant Predictors**:
    - Factors such as systolic and diastolic blood pressure were found to be non-significant in the predictive modeling, suggesting that they may not be reliable indicators for this specific population or set of outcomes.
4. **Resource Allocation Insights**:
    - The analysis emphasizes the importance of using data-driven approaches to allocate healthcare resources effectively, focusing on high-risk populations to improve maternal health outcomes.
5. **Recommendations for Action**:
    - Emphasize targeted interventions for younger women and those with elevated blood sugar levels.
    - Utilize predictive modeling insights to train healthcare providers and develop partnerships for better risk detection and management.

Conclusion and Recommendations

Conclusion

The comprehensive analysis of maternal health risk factors through predictive modeling techniques, including logistic regression and decision trees, has provided valuable insights into the dynamics of pregnancy-related risks. Key factors such as age, blood sugar levels, and heart rate were identified as significant predictors of maternal risk, highlighting the importance of early detection and targeted interventions. The logistic regression model's robust performance, indicated by an ROC score of 0.7773, demonstrates the model's efficacy in distinguishing between "At Risk" and "Not At Risk" categories. Furthermore, the decision tree analysis offers an interpretable framework for understanding the interplay of various factors contributing to maternal health risks.

By leveraging these insights, healthcare providers can make informed decisions to improve maternal health outcomes, allocate resources effectively, and prioritize interventions for high-risk populations.

Recommendations

1. **Targeted Interventions**:
    - Focus on younger women, especially those under 20, who exhibit significantly higher risk levels. Develop tailored educational programs and healthcare strategies to address their unique needs.
2. **Blood Sugar Management**:
    - Implement routine screening and management protocols for blood sugar levels during pregnancy. Enhance patient education on dietary and lifestyle modifications to maintain healthy blood sugar levels.
3. **Monitor Vital Signs**:
    - Increase the frequency of monitoring heart rates in pregnant women, particularly in high-risk groups. Establish guidelines for further assessments based on elevated heart rates.
4. **Utilize Predictive Modeling**:
    - Continue exploring and refining predictive models (e.g., decision trees, random forests) to improve risk detection and classification. Adjust classification thresholds based on ROC/precision-recall curves to enhance decision-making accuracy.
5. **Data-Driven Decision Making**:
    - Foster a culture of data-driven decision-making within healthcare settings. Use insights from predictive models to allocate resources efficiently and improve the effectiveness of interventions.
6. **Training and Education**:

- Provide training for healthcare professionals on using predictive modeling insights to enhance patient care. Incorporate model findings into continuing education programs to ensure staff are equipped with the latest knowledge on risk factors and interventions.

7. **Collaborative Partnerships**:
    - Collaborate with healthcare providers and organizations to facilitate screenings and interventions based on identified risk factors. Establish community outreach programs to raise awareness about maternal health and available resources.

By implementing these recommendations, healthcare providers can significantly enhance maternal health outcomes, leading to healthier pregnancies and reduced maternal mortality rates.