Early Identification of High-Risk Pregnancies: EDA Insights

Problem Statement

In most Primary Health Centers available in the country, especially in the rural areas, Midwives are often unable to identify some danger signs early, so as to be able to tell if a pregnancy is high-risk. So, this project explores maternal health data, to uncover trends and identify various factors that can contribute to a pregnancy being high-risk.

Dataset

The dataset was sourced from a release by Daffodil International University, Bangladesh, and contains anonymized maternal health records. It consists of 1205 records and 12 columns of data. The dataset include both numerical columns (i.e., Age, Systolic BP, Diastolic BP, Blood Sugar (BS), Body Temperature

Objective

- Use exploratory data analysis (EDA) to discover patterns and identify factors that may influence pregnancy risk levels
- Have an idea of the patterns our machine learning (ML) model is likely to follow in its predictions

Data Cleaning

- Null values which were less than 10% of the entire dataset were removed
- Invalid values like Age 325 and BMI 0 were removed
- Outliers were present, but were not removed as they are clinically valid

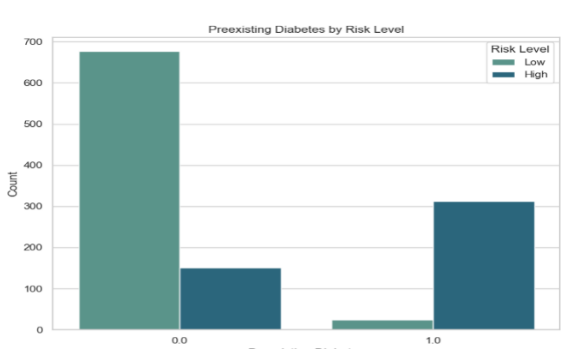
Key Questions Asked

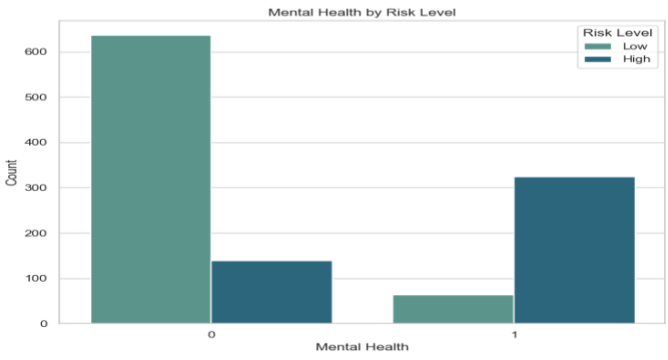
- Which of the variables are more associated with high-risk pregnancies?

Visual Insights

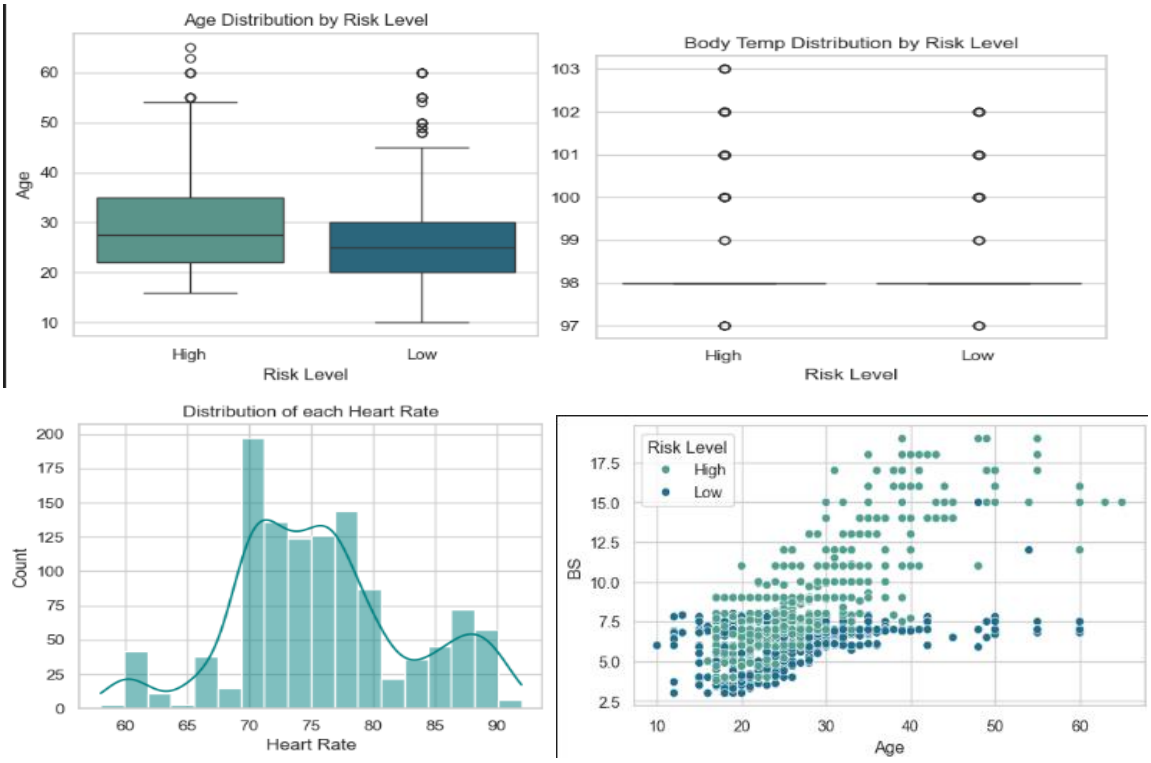
Below are some of the insights gotten from visually analyzing the data

- Most of the pregnant women with pre-existing diabetes and mental health issues were classified as high-risk





- Older Age, High BMI, Blood Sugar level, Systolic and Diastolic blood pressure levels are shown to be associated with high-risk pregnancies. Both high and low-risk pregnancies had high temperatures of higher than 100F. Almost all recorded heart rate in the dataset were all in the normal range for pregnant women (60-90bpm). Example charts can be seen below



Statistical Tests

Statistical tests like T-tests and Chi-square tests were computed to further test and prove that the evidences of relationship between each feature and the target variable, gotten through our data visualization.

- To check whether there is an association between each categorical feature and the target variable, Chi-square Test of Independence was used. The Null Hypothesis (H0) was proposed that there is no association between those variables.

Feature	Chi-Square Value	P-Value	Interpretation
Previous Complications	341.20	0.3×10^{-75}	There is a strong association between the variables and H0 is rejected
Preexisting Diabetes	553.02	$< 0.1 \times 10^{-80}$	
Gestational Diabetes	231.42	0.3×10^{-51}	
Mental Health	462.38	$< 0.1 \times 10^{-80}$	

- For the numerical features, T-test was used to compare their means for each class of risk level. At first glance, there isn't much difference between the means, but statistical tests showed there were differences in their means indicating statistical significance. The table below shows the mean of the features for each group and the statistical values from the test

Variable	High-Risk mean	Low-Risk mean	T-value	P-value
Age	29.71	26.05	-6.57	0.8×10^{-10}
Systolic BP	123.69	112.39	-9.82	0.2×10^{-20}
Diastolic	83.23	73.36	-11.56	0.1×10^{-26}
BS (Blood Sugar)	9.75	6.05	-21.42	$< 0.1 \times 10^{-50}$
Body Temp.	98.66	98.20	-6.48	0.2×10^{-9}
BMI	25.82	21.73	-18.33	$< 0.1 \times 10^{-50}$
Heart Rate	80.07	72.98	-16.33	0.9×10^{-49}

- The p-values, less than 0.05 and t-values that are very far from zero, together with the negative sign which tells the direction that the mean of the first group (High Risk) is greater than the mean of the second group (Low Risk), altogether shows that the difference in mean didn't just happen by chance and that there is a significant difference between the values.