MAIS 202 Deliverable 1
Team 10 - Rachel Ruddy, Lilly Gao, Maya Novichenok, and Sierra Smith

1. **Choice of dataset**:

We chose this dataset because it includes key maternal health indicators like age, vital signs, BMI, and blood glucose levels, which are crucial for predicting high-risk pregnancies. Using a classification model, we aim to enable early risk detection for timely interventions.
Sources:
https://www.kaggle.com/code/yasserhessein/classification-maternal-health-5-algorithms-ml/input
https://www.kaggle.com/datasets/iamsouravbanerjee/maternal-mortality-dataset/data

2. **Methodology**:

a. **Data Preprocessing**:

This data is feasible for a project predicting maternal health because it has structured medical data related to health and pregnancy and all the features being measurable.

The information that is most useful would be the risk level as this is what we are trying to classify, along with systolic and diastolic blood pressure, blood glucose levels, heart rate, and age since these features can all have significant impact on maternal health.

To preprocess the data set we must take several steps to ensure data quality. First, we must remove any missing values either by mean imputation or by removing the entire data entry. Then, we can check for outliers in variables with medical data like blood pressure or glucose levels as if any are too extreme and out of a typical range this could indicate that the data is erroneous and would need to be removed. Finally, we will perform data normalization (rescaling values to be between 0 and 1) and standardization (transforms data to have a mean of 0 and a standard deviation of 1) for the blood pressure, glucose levels, and heart rate data.

b. **Machine learning model**:

We plan to predict maternal health risks using inputs of user data, such as age, BMI, and where the user lives. The machine learning model we propose will have different components. First, we plan to use exploratory data analysis (EDA) to examine the way that different variables relate to certain outcomes. In this case, the variables that we examine are age, location, BMI, blood glucose levels, etc. and we plan to see how these factors correlate with maternal health risks, such as gestational hypertension, preterm labor risks, and gestational diabetes. EDA will help us transform the input data to see which factors are important, which can then be used in the next component of our model, using logistic regression.

We will use logistic regression in order to solve the classification of risk. This method will take the input data and produce an output which indicates the medical risk of a person. Currently, ourn output would be a binary classification, where a person can be either low risk or high risk. However, as we continue exploring how these models work, we plan on deploying an output with more variety (more detailed descriptions of associated risks, and explanation of specific risks related to certain inputs).

Another model we thought about is a random forest. We still have to learn more about this model, but potential pros to using it is that it would allow for more variability of relationships between variables, allowing for more accurate outputs compared to logistic regression. Moreover, this would not require data preprocessing, and it is able to block out noise well. However, the results from a random forest may be harder to interpret and might overfit the data. As we continue learning about machine learning models and talk to our TPM, we plan to refine our machine learning model.

c. **Evaluation Metric**:

To evaluate our maternal health classification model, we can apply several evaluation metrics. We will use the confusion matrix to help identify errors by showing true and false positives/negatives. Accuracy will measure overall correctness, but may be misleading with imbalanced data so we will also use precision to minimize false positives and recall to reduce false negatives, which will be critical in medical cases. We can also use the F1-score to balance precision and recall, ensuring a well-rounded evaluation. Lastly, we can apply the AUC (area under the curve) of the ROC curve (True Positive Rate vs. False Positive Rate) to measure the model's ability/performance in distinguishing risk levels. Together, these metrics will ensure a reliable and effective model for early risk detection.

3. **Application**:

The user is prompted to input relevant details about themself/their health through a text-based form such as their age, BMI, country, blood pressure/glucose, body temperature, and average resting heart rate.

Upon submitting the form, the user will receive an assessment of their pregnancy-related risk level on a scale of intensities ranging from very low to very high risk.