

Maternal Health Risk Classification:
Leveraging Machine Learning for Enhanced Risk Management

Mary Morkos

DSC 2030: Data Mining with Python

Dr. Rudolph Bedeley

26 April 2024

Table of Contents:

Project Description	2
Data-Sources & Preprocessing	2-5
Model	5-8
Results & Discussion	8-9
Summary	10
Work Cited	11

Project Description:

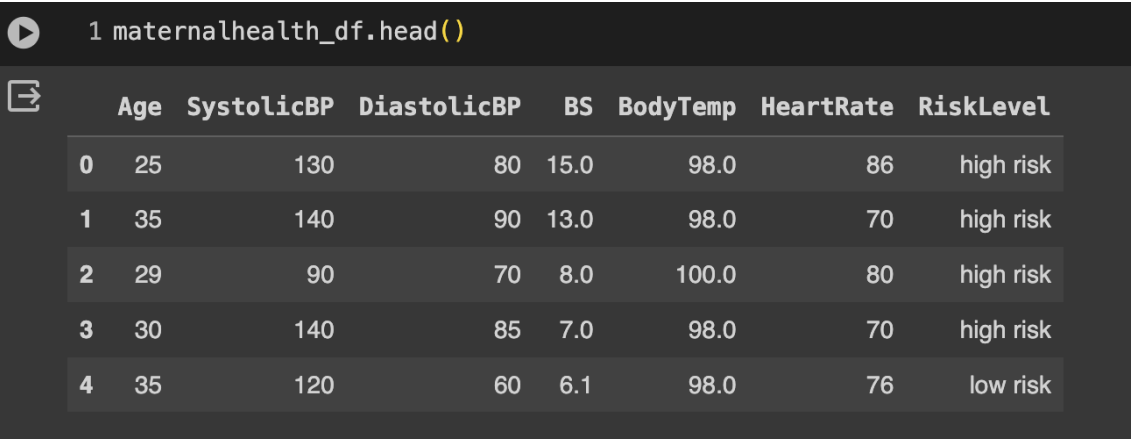
Maternal health is a crucial aspect of public health, with maternal mortality and morbidity being significant concerns worldwide. This project is dedicated to examining maternal health risk data and through the power of machine learning models in order to display an array of techniques of how we can tackle this issue even with our talents as data scientists. To just mention a few statistics in order to comprehend this issue: the World Health Organization (WHO) reports that there are 810 preventable maternal deaths occurring daily worldwide. According to the Centers for Disease Control and Prevention (CDC), approximately 700 women die yearly in the United States due to pregnancy-related complications, which is also due to racial and ethnic disparities in maternal mortality rates. The implementation of risk-based care pathways is of great importance to the unique needs of high-risk patients which can provide us a ray of hope as we continue to understand and battle the existing maternal morbidity and mortality rates. All in all, it is important that we keep the women in our community safe, and we can aim to do that with the skills and talents we have such as occupying data science and leveraging our machine learning skills.

Data-Sources & Preprocessing

To address the research question, I relied on a dataset obtained from Kaggle titled "Maternal Health Risk Data: Predicting Health Risks for Pregnant Patients." This dataset encompasses information gathered from various healthcare facilities, including hospitals, community clinics, and maternal health care centers, through an IoT-based risk monitoring system. The dataset comprises essential attributes crucial for analyzing maternal health risks during pregnancy. These attributes include the woman's age at the time of pregnancy (measured in years), as well as systolic and diastolic blood pressure values (representing the upper and

lower limits in mmHg). Additionally, the dataset includes blood glucose levels expressed in terms of molar concentration (mmol/L), resting heart rate measured in beats per minute, and the predicted risk intensity level during pregnancy based on the aforementioned attributes. Prior to analysis, preprocessing steps were undertaken to ensure data quality and suitability for modeling purposes.

Figure 1: Head of the Dataset



	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	25	130	80	15.0	98.0	86	high risk
1	35	140	90	13.0	98.0	70	high risk
2	29	90	70	8.0	100.0	80	high risk
3	30	140	85	7.0	98.0	70	high risk
4	35	120	60	6.1	98.0	76	low risk

To continue, I performed initial data exploration and preprocessing steps to ensure data quality and suitability for modeling. Before any analysis, the dataset undergoes a data cleaning and preprocessing phase. First, the code checks for missing values in the dataset using the `isnull().sum()` function, which provides the count of missing values for each attribute. Any missing numerical values are then filled with the mean of their respective columns using the `fillna()` method from NumPy. This process ensures that the dataset is complete and ready for analysis without sacrificing significant amounts of data. Additionally, the code checks for duplicate rows in the dataset using the `duplicated().sum()` function, which counts the number of duplicate rows. Duplicate rows can skew analysis results, so it's essential to identify and remove them if present. However, in this case, it seems that there are no duplicate rows, indicating that each entry in the dataset is unique. Additionally, standardization of numerical features and label

encoding of the target variable were conducted to facilitate model training and evaluation. Through this process we can ensure that our code is tidy for further observation and modeling. Below are a few of graphs to visualize our code through simple plots in order to further understand what was in the dataset:

Figure 2: Boxplot illustrating the relationship between Blood Sugar and Pregnancy Risk Level

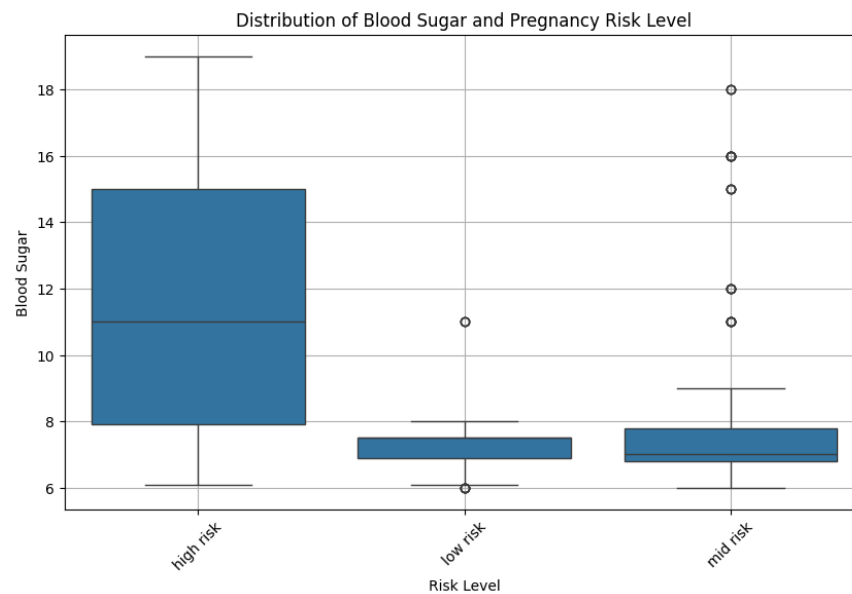


Figure 3: Count distribution of Pregnancy Risk Levels

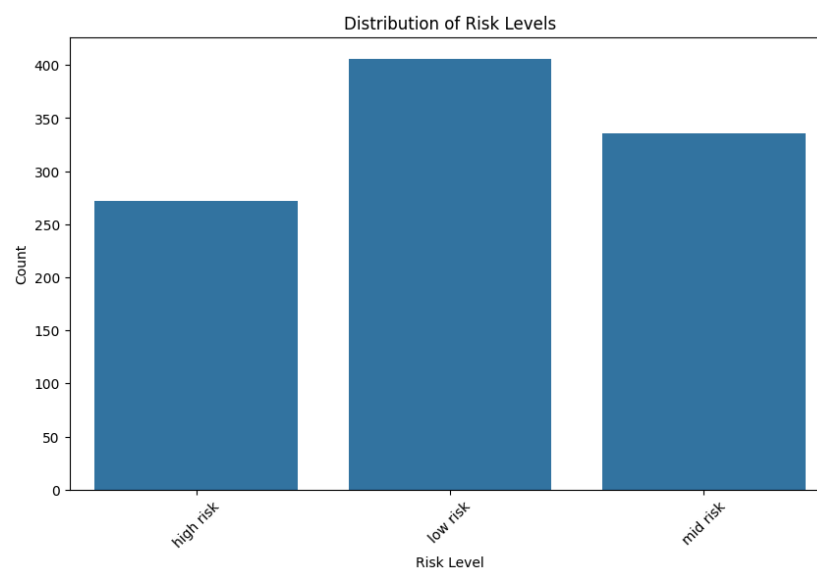
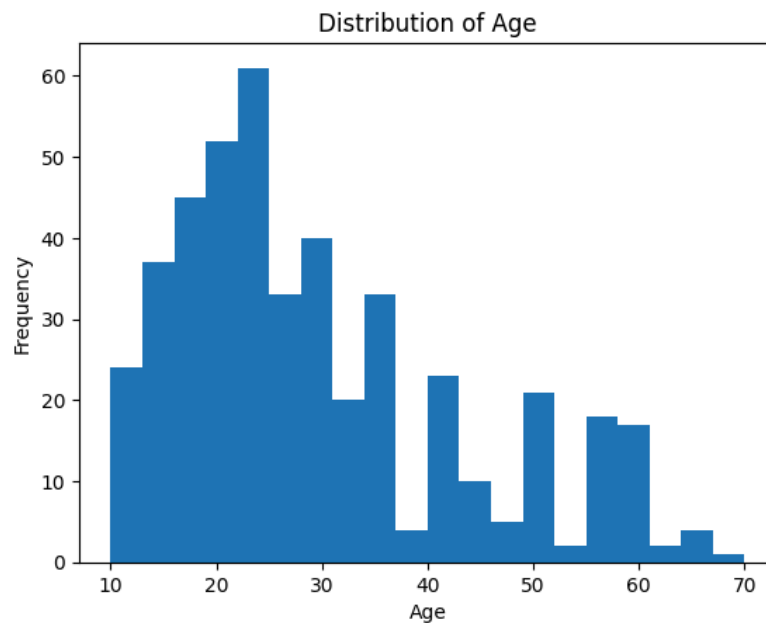


Figure 4: Distribution of Age among Maternal Health Data.



Model

In the exploration of various machine learning algorithms for maternal health risk prediction, I wanted to approach this dataset using a classification technique. The Random Forest Classifier and Linear Regression models were our best performing models. Through the Random Forest Classifier model, evaluation metrics such as accuracy score, it accurately classified risk levels, providing valuable insights into the dataset. Likewise, the Linear Regression is a model to focus on seeing if there is a linear relationship between the feature and target variable through the mean squared error (MSE) and R-squared (R^2) score. Several other models were explored, such as: K-Nearest Neighbors (KNN), Principal Component Analysis (PCA), Logistic Regression, Neural Network model and finally the Decision Tree Classifier.

Furthermore, the analysis employed feature importances and grid search results to optimize model performance and hyperparameter tuning. By visualizing these insights, the code provided a deeper understanding of each algorithm's behavior and effectiveness in maternal

health risk prediction. Overall, the code allowed for us as researchers to grab a better approach to data analysis and predictive modeling in the context of maternal health risk assessment. Below are a few visualizations:

Figure 5: Visualization of feature importances in the predictive model

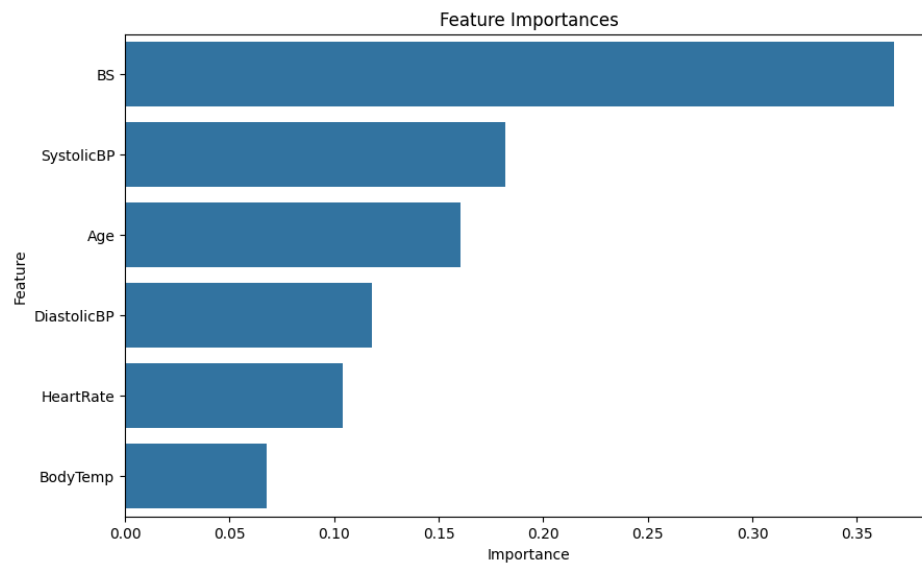


Figure 6: Distribution of Blood Sugar (BS) in Maternal Health Data

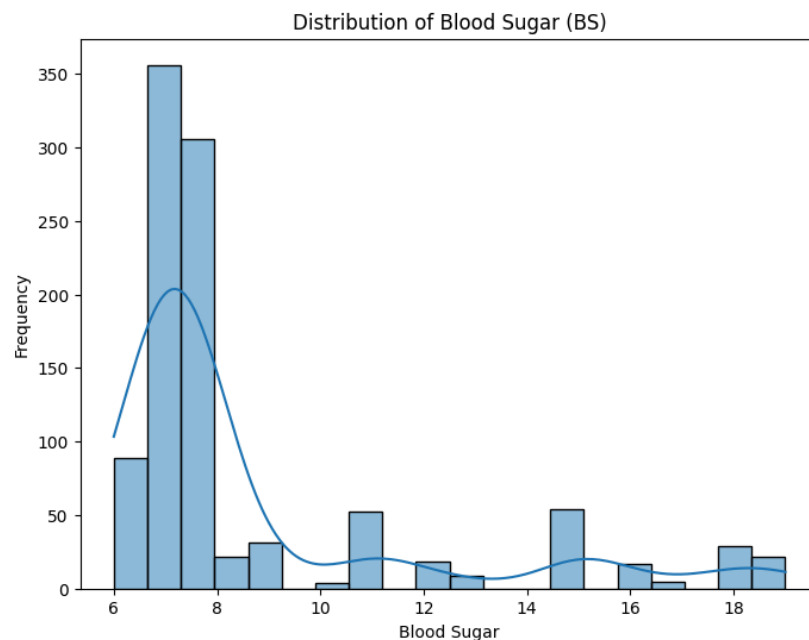


Figure 7: Scatter plot showing the relationship between actual and predicted blood sugar levels.: Mean Squared Error (MSE): 7.811, R-squared (R^2) Score: 0.220

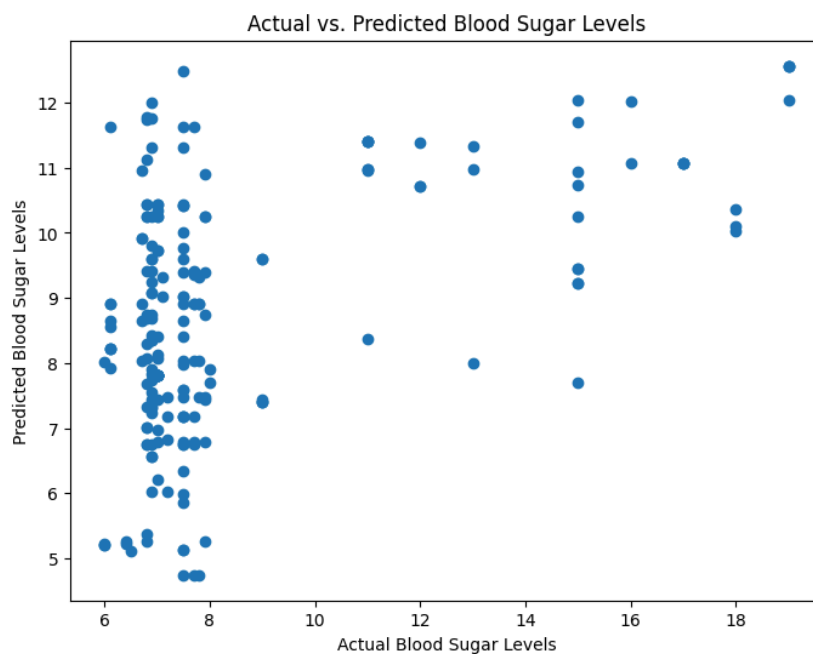
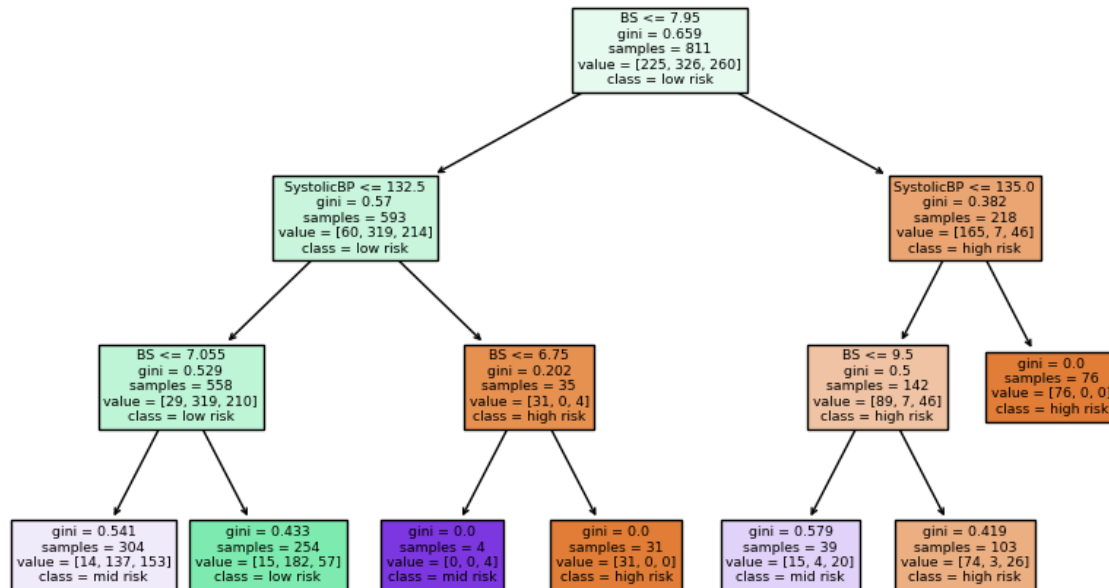


Figure 7: Grid search results visualized as a heatmap showing mean test scores for different combinations of $n_estimators$ and max_depth hyperparameters



Figure 8: Decision Tree Classifier with a maximum depth of 2 trained on the dataset, followed by visualization of the decision tree structure.: Accuracy: 0.650



Results and Discussion

Upon running our machine learning models on the maternal health risk dataset, we got an array of results for each of the models tested. The Random Forest Classifier responded with an accuracy score of approximately 0.84, this meant that this model had great predictivity skills. The Linear Regression model achieved a mean squared error (MSE) of Accuracy: 0.626 and an R-squared (R^2) score of 0.22. Although these are not the most powerful numbers of prediction, it still signifies its effectiveness in explaining the variance in maternal health risk levels. In fact, through our linear regression model we were able to look at insights into the relationship between predictor variables and blood sugar levels during pregnancy. Age, systolic and diastolic blood pressure, body temperature, and heart rate were all examined as potential factors influencing blood sugar levels. The analysis revealed that age had a modest effect, with a

one-unit increase associated with an approximately 0.085-unit rise in blood sugar levels. Similarly, both systolic and diastolic blood pressure showed moderate impacts, with approximately 0.025 and 0.050-unit increases, respectively. Interestingly, body temperature was also a significant predictor, demonstrating a substantial influence on blood sugar levels, with a one-unit increase corresponding to approximately 0.160-unit rise. Moreover, heart rate also had an effect, contributing to a moderate increase of around 0.062 units per one-unit increment. These findings allow us to see the interplay between physiological factors and blood sugar regulation during pregnancy, providing valuable insights for maternal health risk assessment and management strategies.

In addition, these findings emphasize the need for a more comprehensive approach that considers a broader array of factors to enhance predictive performance. Factors such as dietary habits, gestational age, hormonal fluctuations, and genetic predispositions could potentially influence blood sugar levels during pregnancy but were not included in our analysis. Therefore, future research endeavors should aim to incorporate these additional variables to improve the predictive accuracy of models for maternal health risk assessment.

Looking ahead, our models hold significant potential for practical applications in maternal healthcare decision-making. Healthcare providers can leverage these predictive models to allocate resources effectively, prioritize interventions, and improve maternal health outcomes. Furthermore, future iterations of the model could incorporate additional variables and adapt to evolving healthcare trends, enhancing its predictive accuracy and utility. Overall, our analysis underscores the importance of leveraging machine learning techniques in maternal health risk assessment and intervention planning.

Summary

This research delved into predictive factors and modeling techniques for assessing maternal health risks during pregnancy, utilizing the Maternal Health Risk dataset. Analysis revealed the importance of factors such as Age, Blood Pressure (Systolic and Diastolic), Body Temperature, and Heart Rate in predicting pregnancy risk levels. Employing machine learning algorithms like Random Forest Classifier, Logistic Regression, and Linear Regression yielded accuracies ranging from: [0.812, 0.625, 0.648], facilitating the identification of high-risk pregnancies for timely intervention. Blood Sugar (BS) was a critical predictor, and it has great significance in maternal health risk assessment. Despite achieving a Root Mean Squared Error (RMSE) of 7.811 in Linear Regression models for Blood Sugar (BS) prediction, further optimization is essential to improve accuracy, given the range of BS values in the dataset (6.1 to 15.0). The urgency to refine predictive models for maternal health is a global matter and there is a great need for collaborative efforts across diverse academic disciplines. Through international collaboration, we can have a melting pot of expertise and figure out together this great issue through refining: targeted interventions, evidence-based policies, and equitable access to maternal healthcare, which will safeguard the lives and well-being of mothers and children on a global scale.

Works Cited

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 A. Kasruddin Nasir et al. (Eds.), InECCE2019. Lecture Notes in Electrical Engineering, Vol. 632. Springer, Singapore.

Amore, A. D., Britt, A., Arconada Alvarez, S. J., & Greenleaf, M. N. (2023). A Web-Based Intervention to Address Risk Factors for Maternal Morbidity and Mortality (MAMA LOVE): Development and Evaluation Study. *JMIR pediatrics and parenting*, 6, e44615. <https://doi.org/10.2196/44615>

Centers for Disease Control and Prevention. (2019, September 5). *Racial/ethnic disparities in pregnancy-related deaths - United States, 2007–2016*. Centers for Disease Control and Prevention. <https://www.cdc.gov/mmwr/volumes/68/wr/mm6835a3.htm>

CSAFRIT2. (n.d.). Maternal Health Risk Data. Retrieved from Kaggle: <https://www.kaggle.com/datasets/csafrit2/maternal-health-risk-data>

World Health Organization. (n.d.). *Maternal Mortality*. World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/maternal-mortality#:~:text=Every%20day%20in%202020%2C%20almost,dropped%20by%20about%2034%25%20worldwide.>

[Under publication in IEEE] IoT based Risk Level Prediction Model for Maternal Health Care in the Context of Bangladesh, STI-2020.