Maternal Health Risk Predictor

Lennon Au-Yeung, Chenyang Wang, Shirley Zhang (Team 14)

2022-12-04

Contents

Summary

This data analysis project was created in fulfillment of the team project requirements for DSCI 522 (Data Science Workflows), for the Master of Data Science program at the University of British Columbia.

Introduction

Maternal mortality is a large risk in lower and lower middle-income countries, with about 810 women dying from preventable pregnancy-related causes each day (WHO, 2019). Often, there is a lack of information about the woman's health during pregnancy, making it difficult to monitor their status and determine whether they may be at risk of complications (Ahmed and Kashem, 2020). A potential solution to this issue is through using the 'Internet of Things (IoT),' or physical sensors which can monitor and report different health metrics of a patient to their health care provider. Medical professionals can then analyze this information to determine whether a patient may be at risk.

For this project, we aim to answer the question:

> **"Can we use data analysis methods to predict the risk level of a patient during pregnancy (low, mid, or high) given a number of metrics describing their health profile?"**

This is an important question to explore given that human resources are low in lower income countries, and non-human dependent classification methods can help provide this information to more individuals. Further-

more, classifying a patient's risk level through data-driven methods may be advantageous over traditional methods which may involve levels of subjectivity.

IoT sensors can collect a diverse range of health metrics, however not all of them may be useful in predicting whether a patient is at risk of adverse health outcomes (Sutton et al. 2020). Thus, we also hope to use data analysis methods to infer (sub-question) whether some metrics may be more important in determining maternal health risk levels than others.

Methods

The R programming language (R Core Team 2019) and the following R packages were used to perform the analysis: knitr (Xie 2014). The code used to perform the analysis and create this report can be found here: https://github.com/UBC-MDS/maternal_health_risk_predictor/blob/main/doc/final_report.md.

Data

Data used in this study was collected between 2018 and 2020, through six hospitals and maternity clinics in rural areas of Bangladesh (Ahmed and Kashem, 2020). Patients wore sensing devices which collected health data such as temperature and heart rate. The risk factor of each patient was determined through following a guideline based on previous research and consultation with medical professionals.

The full data set was sourced from the UCI Machine Learning Repository (Asuncion and Newman 2007), and can be found here. A .csv format of the data can be directly downloaded using this link. The data can be attributed to Marzia Ahmed (Daffodil International University, Dhaka, Bangladesh) and Mohammod Kashem (Dhaka University of Science and Technology, Gazipur, Bangladesh) (Ahmed and Kashem, 2020).

The data set contains six features describing a patient's health profile, including `age`, `SystolicBP` (systolic blood pressure in mmHG), `DiastolicBP` (diastolic blood pressure in mmHG), `BS` (blood glucose levels in mmol/L), `BodyTemp` (body temperature in Fahrenheit), and `HeartRate` (heart rate in beats per minute). There are 1014 instances in total, with each row corresponding to one patient. Finally, the data contains the attribute `RiskLevel`, corresponding to a medical expert's determination of whether the patient is at low, mid, or high risk (Ahmed et al., 2020).

Planned Analysis

Exploratory Data Analysis

- Figure 1 shows the distribution across target classes, as we can see from the bar chart below, there is not a drastic class imbalance in the training data, however, we will still explore whether a balanced class weight will improve our model performance.

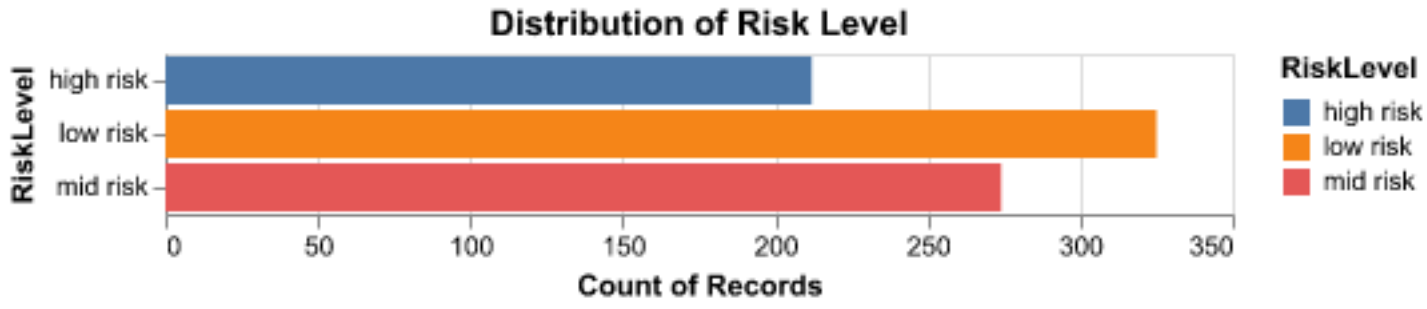

Figure 1: Counts of observation for each class in train data set

Figure 1. Counts of observation for each class in train data set

Table 1: Table 1. Models comparison

Score type	Dummy	Decision Tree	SVM	Logistic Regression	K-Nearest Neighbors
Training Score	0.401	0.928	0.714	0.607	0.797
Mean Cross Validation Score	0.401	0.808	0.695	0.610	0.668

Table 2: Table 2. Confusion Matrix

...1	Predicted High Risk	Predicted Low Risk	Predicted Mid Risk
True High Risk	53	1	6
True Low Risk	1	67	13
True Mid Risk	4	10	48

- Figure 2 shows the density distribution across all features, which could provide us with insights on whether the distribution of some features are different for different target classes.

Figure 2. Distribution of training set predictors for high risk, mid risk and low risk

- Figure 3 shows the features SystolicBP and DiastolicBP have high correlation compared to other pairs of predictors, followed by the correlation between the two blood pressure levels and age. For other pairs of predictors, there are no significant correlations found.

Figure 3. Pairwise relationship between predictors

Results

Model Building

- We have tried the following models: 1. Dummy Classifier; 2. Decision Tree (Myles et al. 2004); 3. Support Vector Machines (SVMs)(Hearst et al. 1998); 4. Logistic Regression; 5. K-Nearest Neighbors (KNN).

For all above models, we used the default parameters and did not include hyperparameter optimization at this stage. Table 1 is the models comparison, and it shows the training scores and mean cross validation scores of the models we tried. Based on the results, we choose Decision Tree model because it has the highest mean cross validation score.

```
## New names:
## * `` -> `...1`
```

- Hyperparameter optimization: For the decision tree model, we use random search method to try different max depth from 1 to 50. From the figure 3, we can see the best depth is 29, and the mean test score is 0.823 which is not bad.
- Table 2 is the confusion matrix which shows the prediction rate is consistently across all risk levels.

```
## New names:
## Rows: 3 Columns: 4
## -- Column specification
## ----------------------------------------------------------------------------------------------------
## (1): ...1 dbl (3): Predicted High Risk, Predicted Low Risk, Predicted Mid Risk
## i Use `spec()` to retrieve the full column specification for this data. i Specify the column types o
## this message.
## * `` -> `...1`
```

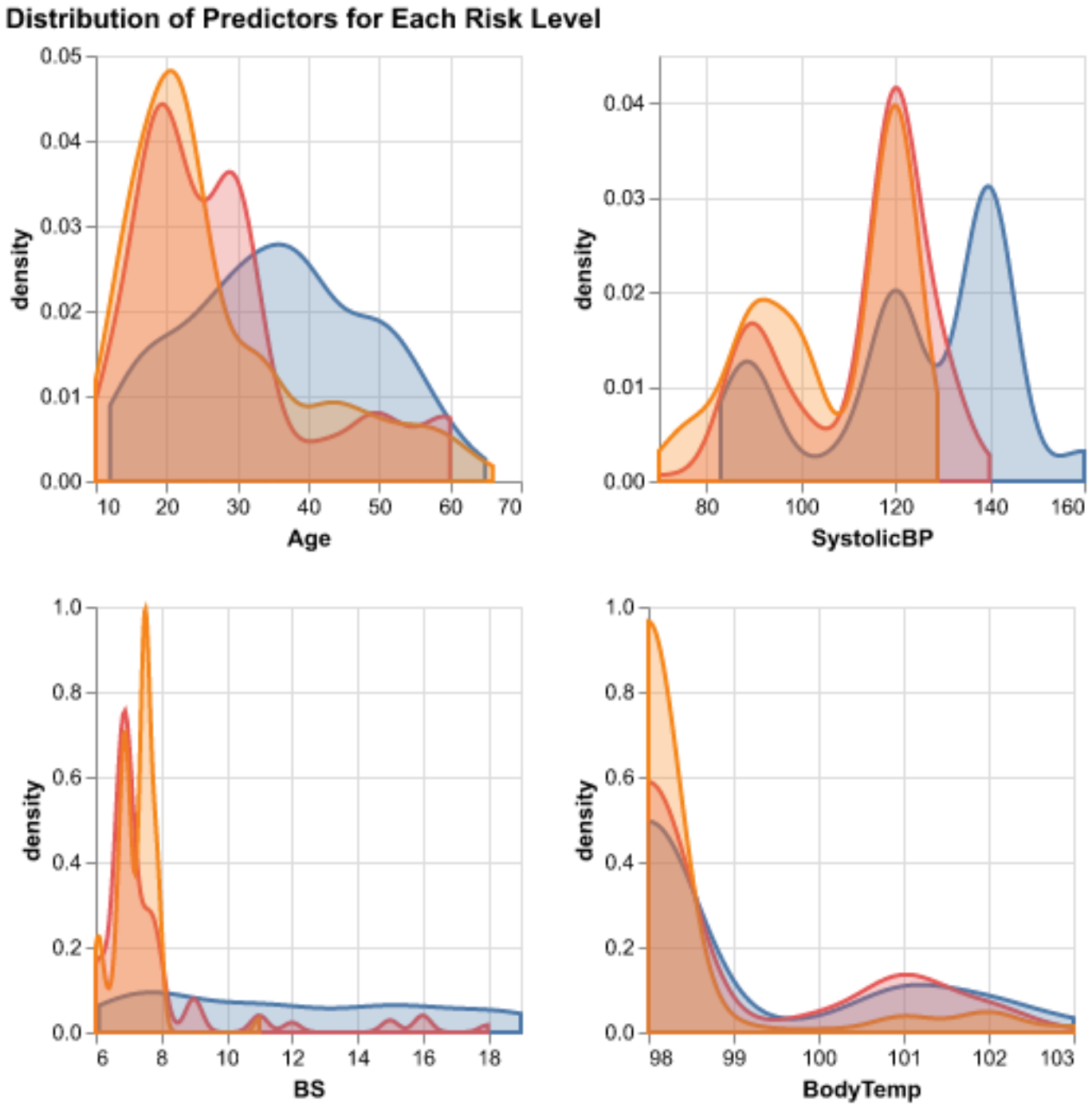

Figure 2: Distribution of training set predictors for high risk, mid risk and low risk

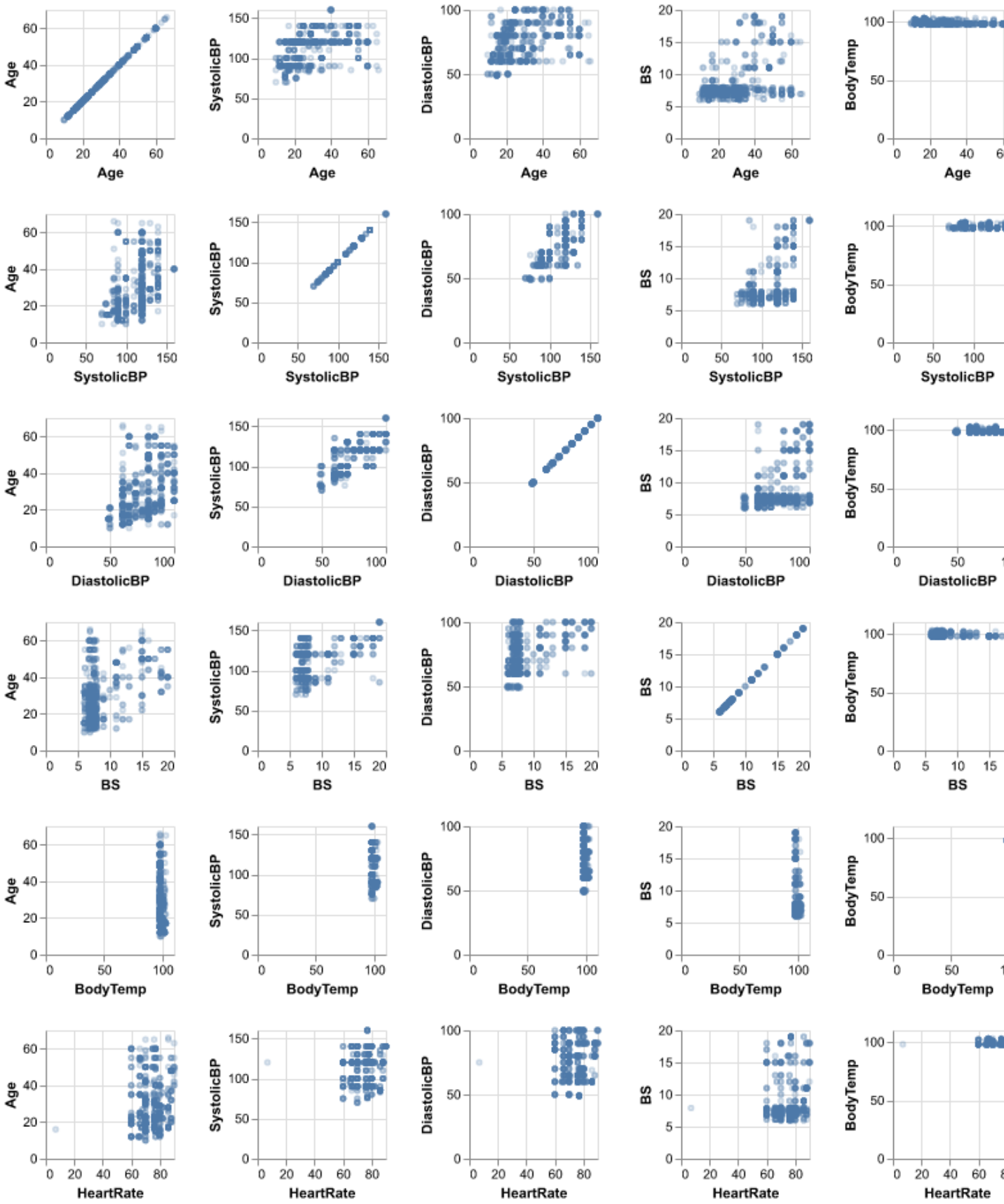

Figure 3: Pairwise relationship between predictors

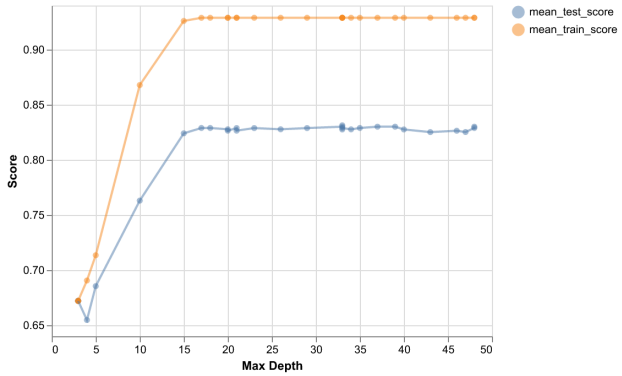

Figure 4: Pairwise relationship between predictors

Assumptions and Limitations

For our analysis, we made the following assumptions: (1) The maternal risk dataset that we used is representative of the population of patients. (2) The risk level classified in the data set is a good indicator of the patient's risk level. (3) The data collected is unbiased.

The dataset we used was collected from the rural areas of Bangladesh, and it might be possible that patients in different regions have different characteristics of health information that affects their maternal risk level, hence the model might not be as accurate when predicting patients from other regions.

Future Directions

As mentioned in the introduction, we are trying to determine whether the patient is at risk, while identifying patients with a high risk level should be our priority. In the future, we could combine low and mid risk level into singular class so that we would have a binary classifier such that we can then explore different classification metrics such as recall instead of using accuracy as our only scoring metrics to evaluate our model. Recall would be an ideal scoring metric in this case as we are trying to minimize the number of false negatives such that high risk patients are not being misclassified by the model.

References

Asuncion, Arthur, and David Newman. 2007. "UCI Machine Learning Repository." Irvine, CA, USA.

Hearst, Marti A., Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. "Support Vector Machines." *IEEE Intelligent Systems and Their Applications* 13 (4): 18–28.

Myles, Anthony J, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. 2004. "An Introduction to Decision Tree Modeling." *Journal of Chemometrics: A Journal of the Chemometrics Society* 18 (6): 275–85.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sutton, Elizabeth F, Sarah C Rogan, Samia Lopa, Danielle Sharbaugh, Matthew F Muldoon, and Janet M Catov. 2020. "Early Pregnancy Blood Pressure Elevations and Risk for Maternal and Neonatal Morbidity." *Obstetrics and Gynecology* 136 (1): 129.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.