Final assignment group 4

Deepika Dilip, Tora Mullings, Daniel Sullivan, Deepa Sharma, Bikram Barua, Newman Okereafor

2022-11-24

Contents

Abstract:

Maternal mortality is a leading public health issue in Bangladesh, with 173 deaths per 100k births. Yet with improvements in public health surveillance, a preventative responsive could be better informed with biomarker data and accurate risk predictions. For this project, we utilize multinomial models to quantify the contribution of biomarkers in predicting mortality risk. We start with multinomial regression, followed by oridinal regression and an xboost model.

Key words:

maternal health, clinical outcomes

Introduction:

Maternal mortality is a leading public health issue in Bangladesh. Advances in public health outreach and medical pipelines have reduced maternal mortality rates, but there remains a glaring gap, especially when considering additional factors, such as socioeconomic status. One of the WHO Sustainable Development Goals was to reduce the global mortality ratio to less than 70 deaths per 100k births

Here, we further explore mortality risk as a product of standard clinical indicators. We obtained this dataset from the UCI repository. Data was aggregated from different sites, including rural and urban health centers.

According to the WHO approximately 810 women die daily due to pregnancy complications (1). With such a high rate of death associated with childbirth it is important to maximize early interventions in high-risk pregnancies in order to monitor and start early intervention to save both the lives of the mother and child. Because of this need, pregnancy has been the focus of many data scientists research in developing many predictive algorithms to try and aid in identifying at risk pregnancies, best emergency interventions and various other aspects to help both mothers and doctors. For this reason, we want to look at identifying low mid and high-risk pregnancies through regression and machine learning methods in order to aid in identifying individuals who could be helped through early intervention.

Literature review:

At risk pregnancies remain a vital research topic despite technological advances and a shrinking pregnancy/childbirth mortality rate. Predictive modeling has been implemented in several ways to reduce pregnancy risks. There are three major groups of studies that have been performed. The largest group predicted risks and complications involved with the pregnancy in specific scenarios (3)() as we are trying to asses with our data set. Many papers also covered predicting delivery methods and successful vaginal delivery (2). The last significant area of study is predicting in vitro fertilization success rates. Our analysis concerns at-risk pregnancies using specific clinical indicators to predict risk. Most studies that predict

complications do it in a much more specific scope. For example, some studies only predict preterm birth or vaginal birth complications while our approach focuses on high-risk births and uses basic vitals. Additionally, our analysis works through generalized linear models and progresses into simple machine learning models, whereas further studies implement more sophisticated models.

Methodology:

Exploratory Data Analysis:

We started by creating exploratory plots to describe positive and negative correlations between predictors and our outcome (risk). We also wanted to make note of the nature of the relationship (e.g. linear or non-linear). Lastly, we used an unsuperivsed approach to classify patients and gain a better understanding of risk heterogeneity.

Data Attributes

- `Age`: Any ages in years when a women during pregnant.
- `SystolicBP`: Upper value of Blood Pressure in mmHg, another significant attribute during pregnancy.
- `DiastolicBP`: Lower value of Blood Pressure in mmHg, another significant attribute during pregnancy.
- `BS`: Blood glucose levels is in terms of a molar concentration, mmol/L.
- `HeartRate`: A normal resting heart rate in beats per minute.
- `Risk Level`: Predicted Risk Intensity Level during pregnancy considering the previous attribute.

Models Used:

Multinomial Regression

First we partitioned the dataset using a 70-30 split. We initially fit a full model with all included variables as predictors. Next, we fit a series of multinomial models, starting with a full model. We then implemented feature selection based on statistical significance to improve accuracy.

XGBoost

We also decided to try using a model that combined previous models with new ones, subsequently increasing accuracy. Therefore, we decided to fit the eXtreme Gradient Boosting algorthim from the `caret` package. In this case, however, we have to split our outcome: one model with predict high risk while the other will predict medium risk.

Ordinal Models

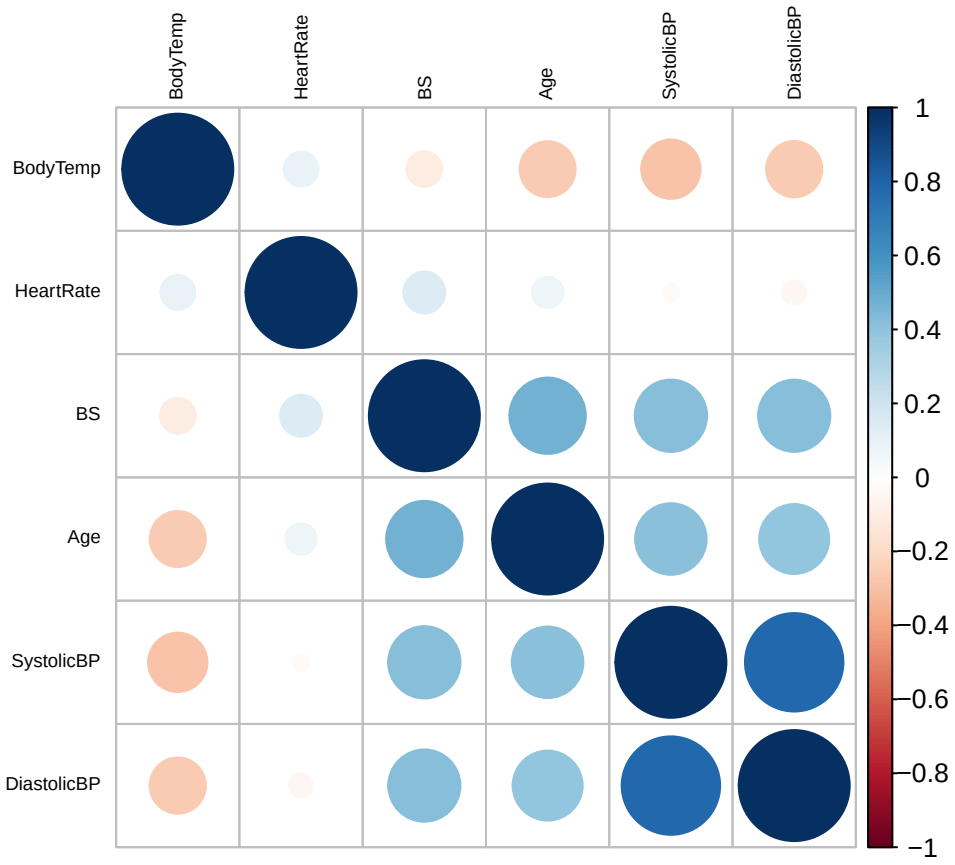
Lastly, we used ordinal models as an alternative to multnomial regression.

Experimentation and Results:

Exploratory Data Analysis

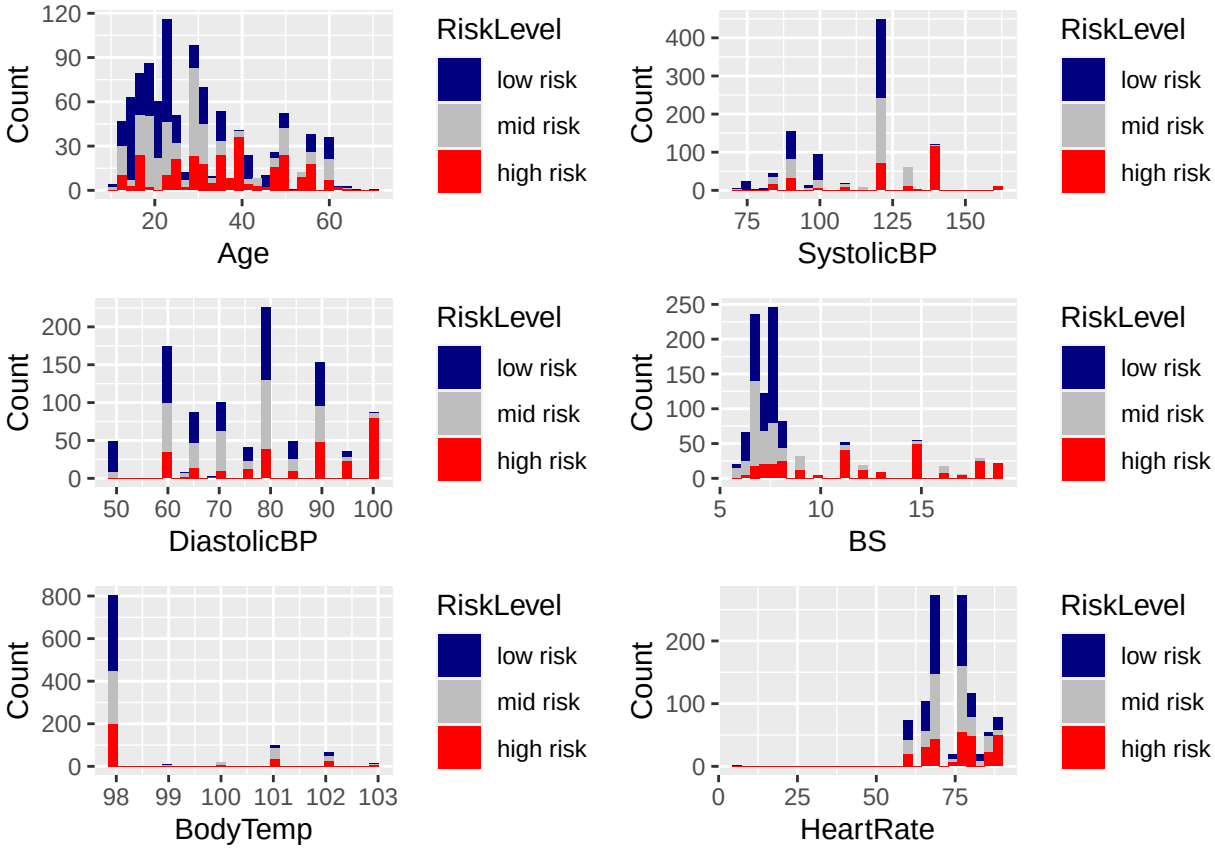
Correlation Plot:

We can start by making a correlation plot to compare continuous values. Age is positively correlated with systolic and diastolic blood pressure.

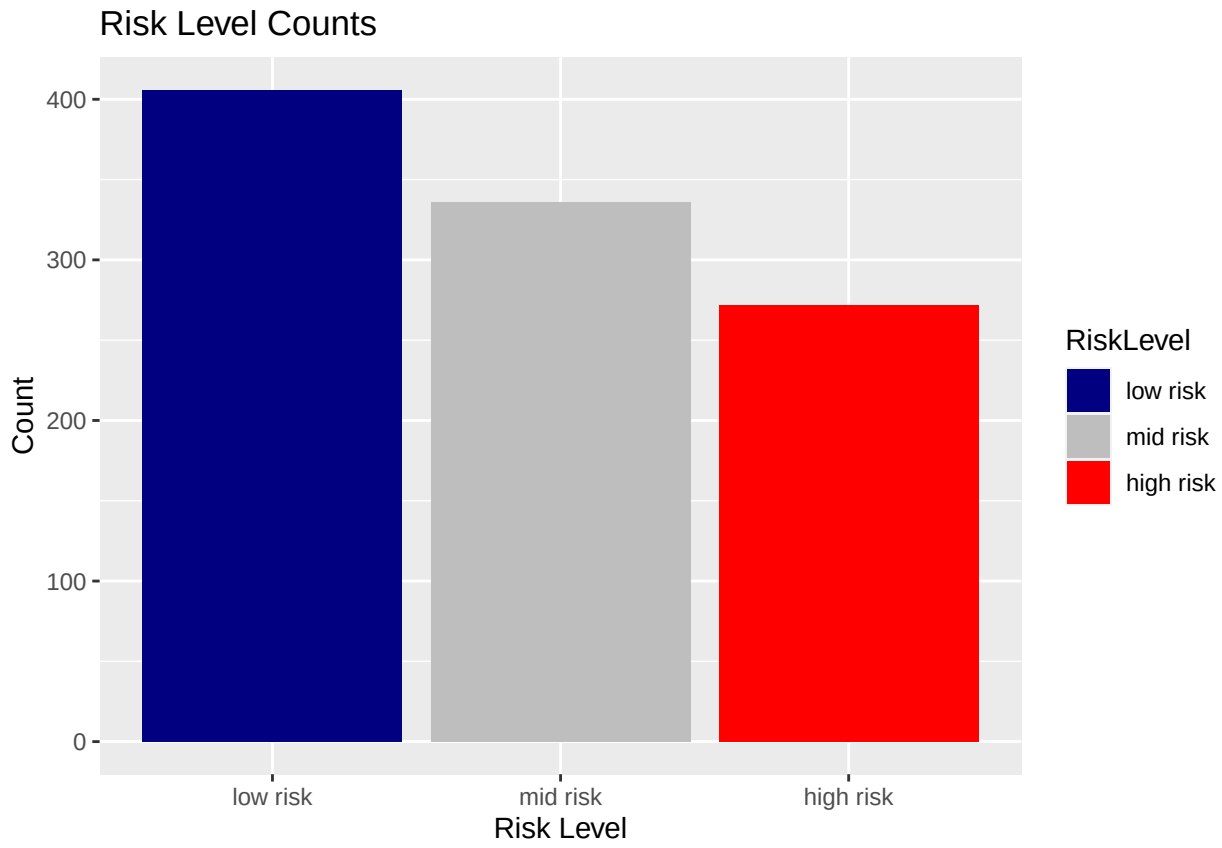


Histograms:

The first step is visualizing data distribution by risk. Here we can see age is skewed, and the extremities of blood pressure and glucose levels are flagged as high risk.


Count
Age
SystolicBP
DiastolicBP
BS
BodyTemp
HeartRate
RiskLevel
low risk
mid risk
high risk

Risk Distribution:



Unsupervised Learning:

One approach we can take is implementing unsupervised clustering since many of the biomarkers are continuous. We can do this by forming a matrix of indicators and seeing if risk levels clusters. From this analysis, we see 6 distinct subgroups: 2 high risk, 3 low risk, and 1 mid risk (with some heterogeneity).

Model Results

Multinomial: Full Model:

metric	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.6356589	0.4062500	0.7125000
Pos Pred Value	0.6307692	0.4020619	0.7307692

Multinomial: Age and Systolic BP as Predictors:

metric	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.6589147	0.3854167	0.6125000
Pos Pred Value	0.5666667	0.3936170	0.8032787

Multinomial: Blood Sugar and Systolic BP as Predictors:

metric	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.7054264	0.2187500	0.6
Pos Pred Value	0.5290698	0.2876712	0.8

Multinomial: Blood Sugar as Predictor:

metric	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.9689922	0.0520833	0.6500
Pos Pred Value	0.5506608	0.3571429	0.8125

Multinomial: model 5 (BS, HR, BodyTemp, systolic BP)

metric	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.6744186	0.3854167	0.6500000
Pos Pred Value	0.6541353	0.3814433	0.6933333

XGBoost Model

Predicting High Risk

	x
Sensitivity	0.9733333
Specificity	0.9875000
Pos Pred Value	0.9954545
Neg Pred Value	0.9294118
Precision	0.9954545
Recall	0.9733333
F1	0.9842697
Prevalence	0.7377049
Detection Rate	0.7180328
Detection Prevalence	0.7213115
Balanced Accuracy	0.9804167

`xBoost`: Predicting Medium Risk

	x
Sensitivity	0.9808612
Specificity	0.9166667
Pos Pred Value	0.9624413
Neg Pred Value	0.9565217
Precision	0.9624413
Recall	0.9808612
F1	0.9715640
Prevalence	0.6852459
Detection Rate	0.6721311
Detection Prevalence	0.6983607
Balanced Accuracy	0.9487640

Ordinal model 1 (all variables)

metric	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.620155	0.3958333	0.6125000
Pos Pred Value	0.640000	0.3486239	0.6901408

Ordinal model 2 (blood sugar, systolic blood pressure, age)

metric	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.6356589	0.4375000	0.5875000
Pos Pred Value	0.5774648	0.4038462	0.7966102

Ordinal model 3 (blood sugar, systolic blood pressure)

metric	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.6201550	0.3958333	0.5875000
Pos Pred Value	0.5594406	0.3689320	0.7966102

Ordinal model 4 (blood sugar)

metric	Class: low risk	Class: mid risk	Class: high risk
Sensitivity	0.8914729	0.15625	0.6375000
Pos Pred Value	0.5693069	0.37500	0.8095238

Summary of Results:

model	accuracy	AIC
mm_model1	0.5836066	1109.227
mm_model2	0.5606557	1208.700
mm_model3	0.5245902	1214.164
mm_model4	0.5967213	1261.509
ordinal1	0.5475410	NA
ordinal2	0.5606557	NA
ordinal3	0.5409836	NA
ordinal4	0.5934426	NA

GLM model construction:

We first tried to predict the level of risk involved with pregnancy using multinomial and ordinal regression modeling. Starting with all variables and then paring down based on correlation to risk level and co-linearity in predictors. We selected first age, systolic blood pressure, and blood sugar level as our first set of restricted predictors, followed by blood sugar and systolic blood pressure for our second. Although we expected to improve the prediction rate by only keeping the most correlated predictors, both model 2(systolicBP, Age, blood sugar) and model 3(systolicBP, blood sugar) in both multinomial and ordinal regressions showed a decrease in predictive accuracy when compared to models with all predictors. The only model in which we saw the best improvement in our predictions was when we only included blood sugar as a predictor. Because assessing pregnancy risk is a critical topic and could lead to life-saving interventions, we want to strengthen this model and suggest a more in-depth analysis of this data via unsupervised learning methods.

Discussion and Conclusions:

Model Interpretation:

First, we look into the multinomial model assessing the accuracy of all predictors and pairing things down to try and improve the model. Starting with every variable, the multinomial model gives an error rate of 41%, which could be better. From here, we eliminated redundancy and low correlation values to the class level. The first pared-down model consisting of Age, Systolic blood pressure, and Blood sugar showed an even worse error rate of around 44%. A pared-down model with blood sugar and systolic blood pressure had a misclassification rate of about 48%. The only improvement was upon making a model solely with blood sugar when the error rate dropped to 40%. This shows that the multinomial regression clearly does not reflect or predict the data too well. One potential future for this type of model would be a boosted multinomial model to help improve the accuracy of predictions.

References:

1 Trends in maternal mortality 2000 to 2017: estimates by who, unicef, unfpa, world bank group and the united nations population division. https://www.unfpa.org/featured-publication/trends-maternal-mortality-2000-2017. Accessed 10 Jan 2021.

2 Birara M, Gebrehiwot Y. Factors associated with success of vaginal birth after one caesarean section (vbac) at three teaching hospitals in addis ababa, ethiopia: a case control study. BMC Pregnancy Childbirth. 2013;13(1):1–6.

3 Gao C, Osmundson S, Edwards DRV, Jackson GP, Malin BA, Chen Y. Deep learning predicts extreme preterm birth from electronic health records. J Biomed Inform. 2019;100:103334.

4 Islam, Muhammed N, Mustafina, Sumaiya N, Mahmud, Tahsin, Khan, Nafiz I Machine learning to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda

Appendices:

Supplemental tables and/or figures.

R statistical programming code.