Machine Learning Foundations - Capstone Project

MACHINE LEARNING MODEL TO PREDICT MATERNAL HEALTH RISK LEVEL

Prepared By: Inuri Muthukumarana

Registration Number: 253

Date of Submission: 30th April 2022

Table of Contents

1 Introduction

Women face many health risks and die due to complications during pregnancy and childbirth. High blood pressure during pregnancy, severe bleeding and infections after child birth, complications caused by chronic heart conditions and diabetes are some of the leading causes of maternal death. Maternal mortality is unacceptably high, especially in rural areas due to lack of information and quality health services. According to WHO, approximately 810 women die every day from preventable causes related to pregnancy and childbirth, and 94% of all deaths occur in low and lower-middle income countries. Improving maternal health and reducing mortality rate of pregnant women is one of the primary concerns of the United Nations SDG.

Age, blood pressure, body temperature and blood glucose levels are some of the identified risk factors accountable for maternal mortality and these health parameters can be easily measured in primary care facilities. More human effort and time is required to manually analyze the health data of pregnant women to identify risk intensity levels. Therefore, a machine learning model that can predict the maternal risk level is essential for early identification of potential health risk in order to come up with comprehensive intervention strategies to minimize maternal deaths and pregnancy-related complications.

Primary objective of this project to build an accurate ML model for predicting maternal health risk levels by comparing several models.

2 Data

Free data set from UCI machine learning repository was used for this project (Link: https://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set). It was provided by a research group from Dhaka University of Science and Technology, Daffodil International University and University Malaysia Pahang. The dataset is in .csv file format (30 KB) and contains data collected from a IoT-based health monitoring system from various hospitals, clinics and maternity care centers in rural areas of Bangladesh.

The data set consists of 7 attributes (columns): age in years, systolic blood pressure in mmHg, diastolic blood pressure in mmHg, blood sugar in mmol/L, body temperature in F, heart rate in beats per minute, risk intensity level (Low, Mid and High). Out of all 1014 data entries, there are 406 records of low risk, 336 records of mid risk and 272 records of high risk.

3 Methodology

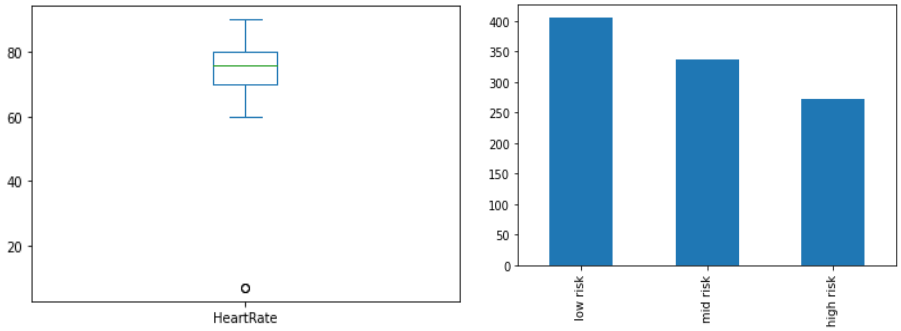
Since there three risk classes as low (0), mid (1) and high (2), multi class classification approach was used for this project. For model development, python libraries such as numpy, pandas, seaborn, matplotlib, sklearn were used.

3.1 Pre-Processing and Exploratory Data Analysis

Before model building, exploratory data analysis and several data pre-processing steps were performed to get a better understanding of the available data variables, clean and transform data.

i. Check data types and descriptive statistics to identify the characteristics of the data set

ii. Graphs (Bar charts, Boxplots)

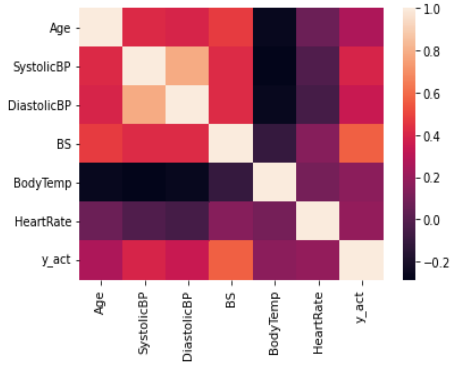


iii. Treat missing values and outliers in the data set

iv. Encode “RiskLevel” categorical column to numeric

3.2 Model Building

i. Based on the correlation matrix, Independent variables (X - features of the model) and dependent variable (y) were decided.




```
[ ]  X_variables = ['Age', 'SystolicBP', 'DiastolicBP', 'BS', 'BodyTemp', 'HeartRate']
     y_varibale = 'y_act'
```

ii. The data was divided into train and test data in the ratio of 80:20.

iii. Four multiclass classification models; Decision Trees, Random Forest (RF), Support Vector Machines and k-Nearest Neighbors were built and tested, to select the most suitable model.

3.3 Model Comparison and Selection

The performance of each model was evaluated using Accuracy, Precision, Recall and F1 score. Based on the scores, RF classifier was selected as the best classifier for this problem.

Model	accuracy	precision	recall	f1_score
Decision Tree	0.753695	0.776414	0.753695	0.752239
Random Forest	0.832512	0.838724	0.832512	0.832419
SVM	0.650246	0.635250	0.650246	0.626198
K-Neighbour	0.689655	0.689837	0.689655	0.688362

3.4 Hyper parameter Tuning

The hyper parameters of RF classifier were tuned using "Randomized Search" method to get best results and final model was saved as a 'Pickle' file.

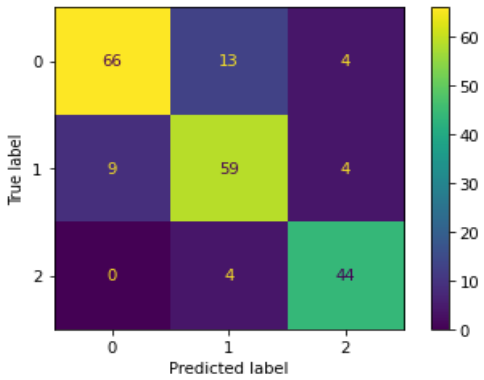
3.5 Model Evaluation

Evaluation matrices for the final RF model were calculated.

4 Results

The confusion matrix and evaluation matrices of the model are as follows.

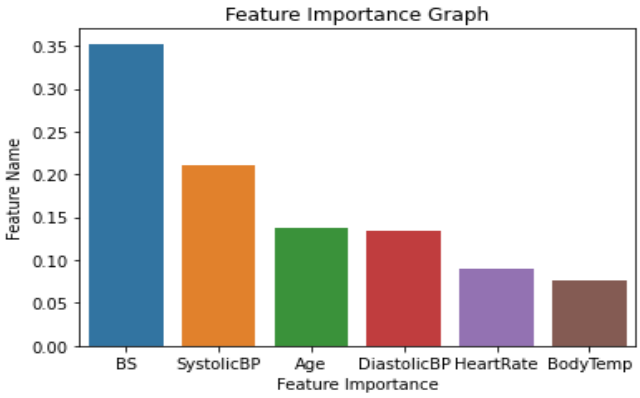
```
f1_score: 0.8324491505285918

classification_report:
              precision    recall  f1-score   support

           0       0.88      0.80      0.84        83
           1       0.78      0.82      0.80        72
           2       0.85      0.92      0.88        48

    accuracy                           0.83       203
   macro avg       0.83      0.84      0.84       203
weighted avg       0.84      0.83      0.83       203
```




```
Weighted_auc_score_ovo: 0.9291768881766804
Weighted_auc_score_ovr: 0.9212631143534011
```

The feature importance table is given below.

	feature	importance
3	BS	0.352403
1	SystolicBP	0.210073
0	Age	0.137793
2	DiastolicBP	0.134168
5	HeartRate	0.090036
4	BodyTemp	0.075528



5 Conclusion

Based on the above findings, RF multiclass classifier provides best accuracy, f1 score, recall of 0.83 and precision of 0.84. The weighted average of AUC of each class against the rest is 0.92. According to the feature importance table, blood sugar level is identified as most significant risk factor that can increase the maternal mortality risk. Physicians can use this machine learning model as a supporting tool to accurately determine the maternal health risk level. The model will also be useful for early identification of potential risk factors. Hence, it will help to reduce the burden on health staff and systems.

6 Discussion

In the dataset used, the three classes are not represented equally, causing multi class imbalance. ML techniques should be used in the future to address the class imbalance issue and improve the performance of the classifier. Besides, the above model is trained and tested using limited data collected from rural areas of Bangladesh. It would be better if the model can be trained and tested in different settings and contexts (Eg: urban, suburban, developed, developing and underdeveloped countries) to make it a generic model.

7 References

1. WHO. "Maternal Mortality". World Health Organization, 2019, https://www.who.int/news-room/fact-sheets/detail/maternal-mortality
2. Ahmed, M., 2020. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set. Daffodil International University