A Comparison of Decision Trees and Random Forests on Maternal Health Dataset

Ho Hang Cheng

Description and Motivation

We would build a Decision Trees (DT) model and a Random Forests (RF) model on the Maternal Health dataset to compare the performance of the two selected methods. Both best-trained models would be run on the testing set to make comparisons and analyzations from predicting the Risk Level of pregnant women (low risk, mid risk, high risk). Maternal mortality remains high in many parts of the developing world [1], and this problem has always been underestimated. The aim of the study is to build a reliable model for detecting the risk level of pregnant woman and prevent maternal mortality.

Basic Statistic of the Dataset

- The Maternal Health Dataset is from the UCI Machine Learning Repository.
- The dataset consists of 1014 rows and 7 columns.
- Age, Systolic Blood Pressure (SystolicBP), Diastolic (DiastolicBP), Blood Sugar (BS), Body Temperature (BodyTemp) and Heart Rate are the predictors for the target column: Risk Level. Risk Level is the class label which consists of low risk, mid risk and high risk.
- There is no missing value in the dataset.
- The pie chart (Figure 1) shows the proportion of three risk levels in the dataset. Although the low risk's sector shares the largest proportion, which accounted for 40%, the difference in the number of instances is not significant among the three categorial risks. Mid risk accounted for 33% and high risk accounted for 27%. The class labels are quite balanced.
- The table (Table 1) shows a brief description about the numeric features of the dataset. Values are rounded into 2 decimal places.
- In the correlation heatmap (Figure 2), the labels (Risk Level) change into 1 (low risk), 2 (mid risk) and 3 (high risk) to see which variables have the highest correlation with Risk Level. SystolicBP and BS have relatively high value among the others, it seems they influence the Risk Level most. BodyTemp and Heart Rate are likely not important as the other factors.
- The box plot (Figure 3) shows the distribution and outliers of each attribute which were grouped by Risk Level. From the figure, Age, SystolicBP, DiastolicBP and BS got high values at high risk, and BS has the most significant increase in value compared to itself at low risk and mid risk. BS is likely an important factor.

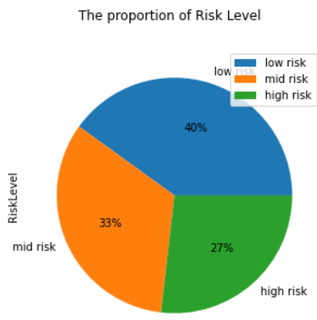

Figure 1. Pie Chart

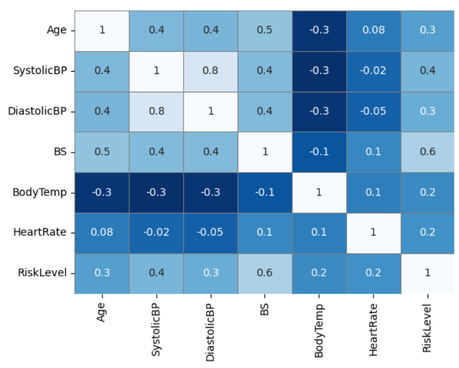

Figure 2. Correlation Heatmap

Attributes	Mean	Std	Min	Max	Q1	Q2	Q3
Age	29.87	13.47	10.00	70.00	19.00	26.00	39.00
SystolicBP	113.20	18.40	70.00	160.00	100.00	120.00	120.00
DiastolicBP	76.46	13.89	49.00	100.00	65.00	80.00	90.00
Blood Sugar	8.73	3.29	6.00	19.00	6.90	7.50	8.00
Body Temp	98.67	1.37	98.00	103.00	98.00	98.00	98.00
Heart Rate	74.30	8.09	7.00	90.00	70.00	76.00	80.00

Table 1. Maternal Datasets description

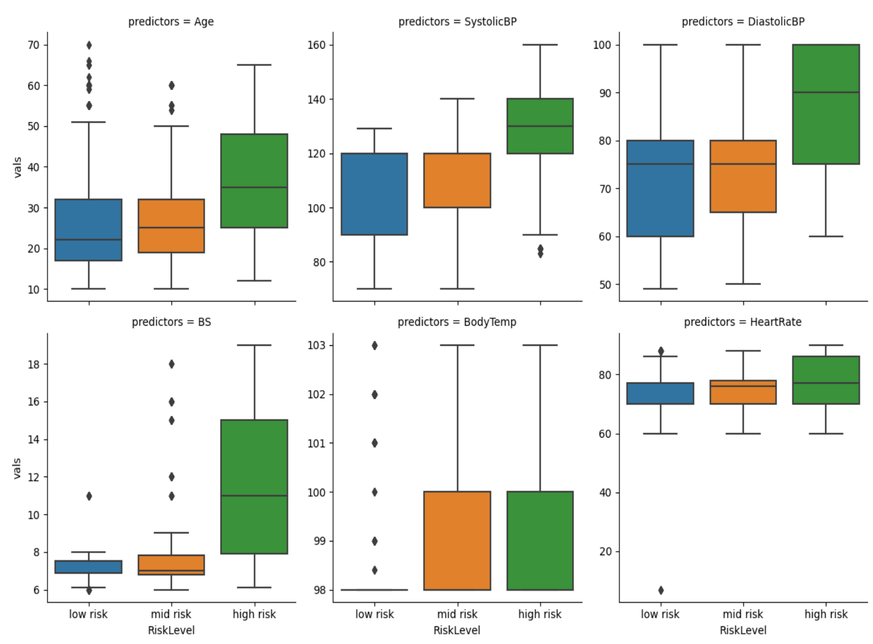

Figure 3. Box plot

Decision Trees (DT)

- A widely used supervised learning method for classification and regression in a tree shape diagram. The selection process can be described as a sequence of binary selections through a tree structure [1]. By learning decision rules inferred from the data structure, the model would predict the value of target variable.
- It starts from a root node which represent a test on an attribute to make the first split decision, then internal nodes would be constructed continually to build a tree like model. At the end, the leaf node at the bottom represent the target labels.

Pros	Cons
➢ Requires less effort for data pre-processing	➢ The unstable model is sensitive to changes in data
➢ Easy to understand and interpret [2]	➢ Can easily overfit. Complex decision rules can be made as it has no inherent mechanism to stop. [3]
➢ Use a white box model	➢ Limited performance in regression [4]

Random Forests (RF)

- It is also a supervised learning method in machine learning for classifications and regressions, which builds a forest with an ensemble of decision trees. Bagging and feature randomness were used to build each decision tree to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree [5].
- Label with the most votes among the trees would be the prediction of RF.

Pros	Cons
➢ Generally, with high accuracy [6]	➢ Not easy to interpret. It doesn't provide visibility into coefficients.
➢ Performs well even work with high number of features in data	➢ Can be computationally intensive for large dataset
➢ Not influenced by outliers to a fair degree	➢ It is like a black box algorithm, can only have very little control over the model. [7]

Hypothesis Statement

- Regarding to initial knowledge on both methods, RF would likely to have a better performance than DT
- RF is expected to have a longer test time , depends on number of trees it would build.

Choice of Parameters and Experimental Result

Decision Trees:

- We used classification tree in CART algorithm to train decision trees. Gini impurity is used to split training set into a decision tree model.
- To select appropriate tree depth, 'Min Leaf Size' is the parameter chosen for optimization, which is the minimum observations each leaf has. We performed a cross-validation on the model and find the best leaf size with minimum cross-validated error. Regarding to the result in figure 4, the best leaf size is 8.
- Figure 5 and 6 are the confusion matrix and the ROC curve of testing model respectively.

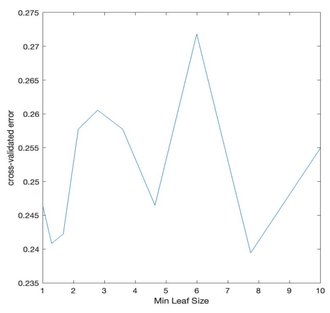

Figure 4. Min Leaf size against cross-validated error

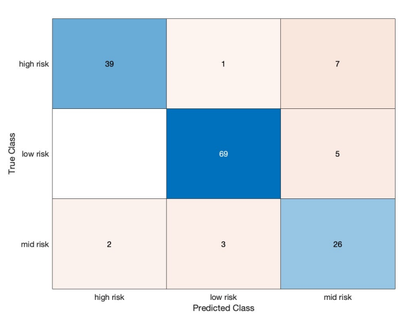

Figure 5. Confusion Matrix (DT)

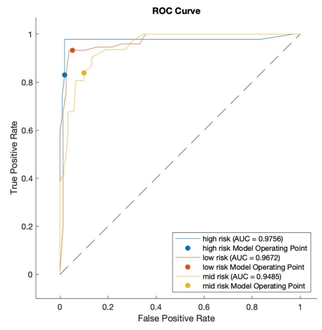

Figure 6. ROC Curve (DT)

Random Forests:

- Use bootstrapping of the training set. Attributes are randomly selected to make trees with randomly selected samples in training set.
- Select the number of ensemble learning cycle, which is the number of tree in forest, as the hyperparameter. We plot the out-of-bag classification error over the number of tree to find the optimal value. From figure7, number of grown tree at 90 to 100 have lowest Out of Bag Error.
- Set number of tree in forest as 95.

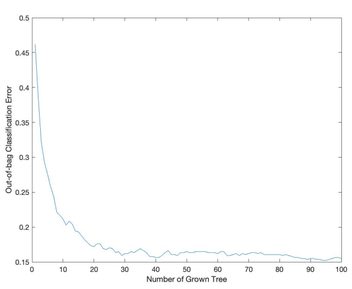

Figure 7. Number of Tree against OOB Error

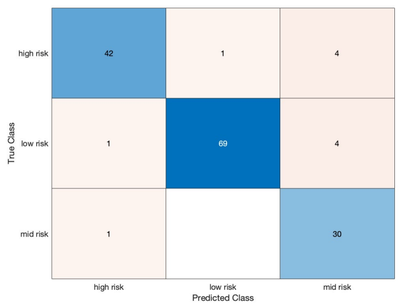

Figure 8. Confusion Matrix (RF)

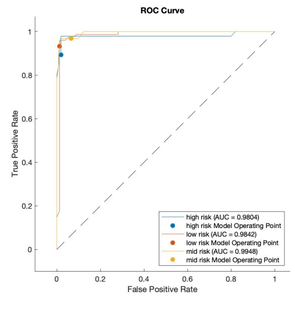

Figure 9. ROC Curve (RF)

Methods	Train Acc.	Val. Acc.	Test Acc.	Precision	Specificity	Recall	F1 Score	Avg. AUC
DT	0.8732	0.8728	0.9211	0.8602	0.9435	0.8670	0.8636	0.9638
RF	0.9531	0.9649	0.9518	0.9099	0.9673	0.9313	0.9205	0.9865

Table 2. Results Table

Methodology

- Splitting the dataset into three parts, which are the training set (70 %), the testing set (15%) and the validation set (15%). Validation set is used to tune the parameter of the models, and testing set would be used to assess the performance of the final models.
- Apply 10-fold cross validation on splitting data and training models to detect overfitting
- Optimize models by tuning the parameters present in the models, until they are best trained.
- Perform testing set on the trained models and make comparison on results by different scoring metrics, for instance the confusion matrices, precision, recall and F1 score.

Analysis and Critical Evaluation of Result

- For the DT model, we performed a 10-fold cross validation to protect against overfitting. From the 10 models we selected one of them with the highest accuracy in testing set as the final model of DT. The accuracy on the testing set is higher than training set by 4.8%, as there would be more noises in the training set due to the proportion of training set (70%) is much larger than testing set (15%) when splitting the original data, and DT model is sensitive to outliers.
- Cross validation didn't apply on RF, because bootstrapping is already a resampling method performed in RF. The out-of-bag error over the number of grown tree decrease as the number of tree increase in figure 7. Thus, the hyperparameter number of trees is set to be 95 to build model with the lowest out-of-bag error. There are no significant difference between accuracy on training set and testing in RF model, it is same in the DT model. Thus, there are no over-fitting in both model.
- Comparing the accuracy in training, validation and testing set, RF have a better performance in overall. For the testing set, RF's accuracy is about 8% higher than DT. It is because RF used bootstrapping which randomly selected variables to build number of trees, and predict the labels by majority voting, in most of the case it would be more accurate than the DT, which only use one tree in the model.
- In case there are class imbalances, precision, recall, specificity and F1 score are calculated from the confusion matrices of the two models (Figure 5 & 8). These scoring metrics interpret the predictive performance of the models. Considering the case of the dataset, the most undesirable scenario in our predictions is that pregnant women with high risk is classified into low risk class, which may result in some pregnant women do not receive adequate support and place two's life in danger. In this case, recall is more suitable than precision for measuring the accuracy of the model, because recall can measure the extent of error caused by the FN cases mentioned above. And with higher Recall and Precision, the higher the F1 score [8]. Thus, we use Recall and F1 score to make comparison on two models. From table 2, RF have a better result in both scoring metrices, about 6 % better than DT in Recall and 5.7% better in F1 score, which means the possibility of False Negative cases happened in RF's model is less than DT 's. Moreover, RF have higher scores in Precision and Specificity as well.
- Figure 6 and 9 are the ROC curve of DT and RF respectively. The curves in both figures are close to top left corner, which indicates performance of both models are good. We summarize ROC curve by the average AUC to make comparison, and RF is slightly better than DT by about 2%. Overall RF is a better machine learning method for predicting label in this case.
- DT have used 0.5869 second to train the model, and RF used 3.1817 seconds, both models are trained in MATLAB. The time for RF needed is about 5 times the DT's.

Lesson Learned and Future Work

Lesson Learned:

- Data pre-processing can affect the result very much, for example data cleaning can enhance the performance of DT model as it is sensitive to outliers, if we remove the noise before building the model, we will have a better DT model with higher accuracy on prediction. It is also important to notice if there is class imbalance in the original dataset, it would affect the performance in classification if one of the class have the majority in the data, there would be higher probability having misclassification.

Future Work:

- Compare with real-world data, the size of dataset in this study is relatively small. The future work of this research is to collect more samples from different sources with balanced classes.
- Apply feature selections and make comparison with initial result
- Try different algorithms in tuning the hyperparameters, tune more hyperparameters.
- Remove the outliers during data pre-processing

References

[1] Ashraf, N., Field, E., Rusconi, G., Voena, A., & Ziparo, R. (2017). Traditional Beliefs and Learning about Maternal Risk in Zambia. *The American Economic Review*, *107*(5), 511–515. http://www.jstor.org/stable/44250451

[2] Attewell, P., Monaghan, D. B., & Kwong, D. (2015). CLASSIFICATION TREES. In *Data Mining for the Social Sciences: An Introduction* (1st ed., pp. 162–184). University of California Press. http://www.jstor.org/stable/10.1525/j.ctt13x1gcg.13

[3] BISHOP, C. H. R. I. S. T. O. P. H. E. R. M. (2016). *Pattern recognition and machine learning*. SPRINGER-VERLAG NEW YORK.

[4] Mallach, E., & Berger, P. D. (1975). Decision Trees with Continuous Distribution. *Operational Research Quarterly (1970-1977)*, *26*(2), 297–304. https://doi.org/10.2307/3008463

[5] Kapil, A. R. (2022, October 1). *Advantages and disadvantages of Decision Tree in machine learning*. Blogs & Updates on Data Science, Business Analytics, AI Machine Learning. Retrieved December 20, 2022, from https://www.analytixlabs.co.in/blog/decision-tree-algorithm/

[6] Scornet, E., Biau, G., & Vert, J.-P. (2015). CONSISTENCY OF RANDOM FORESTS. *The Annals of Statistics*, *43*(4), 1716–1741. http://www.jstor.org/stable/43556658

[7] Singh, J. (2020, December 26). *Random Forest: Pros and cons*. Medium. Retrieved December 20, 2022, from https://medium.datadriveninvestor.com/random-forest-pros-and-cons-c1c42fb64f04

[8] LT, Z. (2022, February 25). *Essential things you need to know about F1-score*. Medium. https://towardsdatascience.com/essential-things-you-need-to-know-about-f1-score-dbd973bf1a3#d3cf