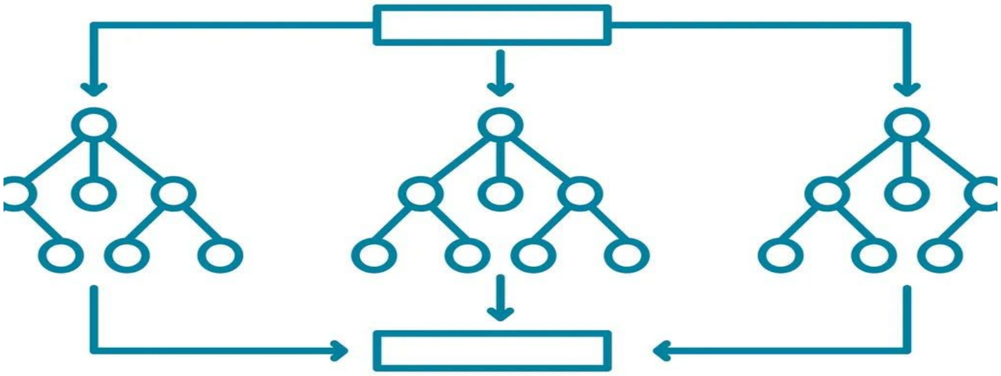
Radboud University

Prediction of Delirium Incidence Using Random Forest Classifier

BMS59 - Prediction Models and Machine Learning



***Authors**: Maxwell Agyemang (s1120644)*
Thomas Bouwmeister (s1018079)

22 – 11- 2024

Motivation

In this paper, we propose a Random Forest (RF) classifier to predict the incidence of delirium using a dataset obtained from Radboudumc's Department of Intensive Care. The dataset is imbalanced since there is an unequal distribution of participants between the 'no delirium' (n=5,473), hypoactive (n=639), hyperactive (n=255), and mixed (n=431) classes. We denote the 'no delirium' class as 'majority' and the other classes as 'minority'. Generally, this imbalance can cause classifiers to become biased toward the majority class, resulting in poor performance for minority classes (1). This leads to misleading accuracy metrics, as the classifier may predominantly predict 'no delirium', failing to identify the minority classes correctly. Consequently, precision and recall may be skewed, giving a false sense of the model's effectiveness and potentially missing patients with rare but significant delirium subtypes. Here, the random forest classifier was employed to address these concerns.

RF is an ensemble method consisting of regression trees, generated from bootstrap samples of the training data (2). By using bootstrap sampling, RF ensures that the minority classes are represented in each subset, thus giving more importance to the minority class. However, similar to most classifiers, RF can also suffer from the curse of learning from an extremely imbalanced training data set (3). RF can result in biased predictions toward the majority class. Again, RF can be computationally intensive, especially with large datasets. While RFs are generally good at reducing overfitting compared to most classifiers, they can still overfit, particularly if the number of trees is too large. To alleviate this problem, RF allows for the incorporation of cost-sensitive learning and balanced sampled data through two processes known as Weighted Random Forest (WRF) and Balanced Random Forest (BRF), respectively (4). By balancing sample data, BRF provides equal representation of classes, improving learning, reducing bias, and enhancing performance metrics. Furthermore, given that RF can adopt cost-sensitive learning, misclassifications of the minority class can be penalized by assigning higher weights to minority classes, thereby improving its performance on the minority class. With the versatility of RF, it achieves high accuracy and is essentially robust to noise and outliers in the data. Additionally, RF provides estimates of feature importance, helping to identify the most influential predictors in the dataset.

In summary, the Random Forest classifier is a good model for this dataset due to its high accuracy, versatility, and robustness. It has the ability to handle complex data, provide feature-importance insights, and reduce overfitting.

Results and Interpretations

From the scree plot of the Principal Components (*in R code*), it was observed that the first principal component (PC1) explained 20% of the variance, while the second and third principal components (PC2 and PC3) each accounted for approximately 6%. The scatter plot of PC1 versus PC2 showed no remarkable differences or clusters between genders or between scheduled and unscheduled patients (Figure 1). Furthermore, exploratory analyses with categorical variables such as CPR, COPD, diabetes, and the future outcome variable 'delirium class' revealed no noteworthy patterns. These variables will therefore not be included as predictors in the final model. The delirium class still will be used as the outcome variable.

After the PCA, we developed an RF model utilizing five features: age, urea concentration, APACHE-II score, type of patient, and infection. When applied to the training data, the model demonstrated an accuracy of 93% and a kappa statistic of 0.76, indicating a high level of agreement between the model's predictions and the actual delirium classifications (Table 3). The RF feature importance analysis revealed that urea concentration was the most significant predictor, followed by age and APACHE-II score. These features were the primary contributors to the model's predictive performance, whereas the type of patient and infection were less influential and contributed minimally to the predicted outcomes. Overall, The algorithm performance is fairly decent.

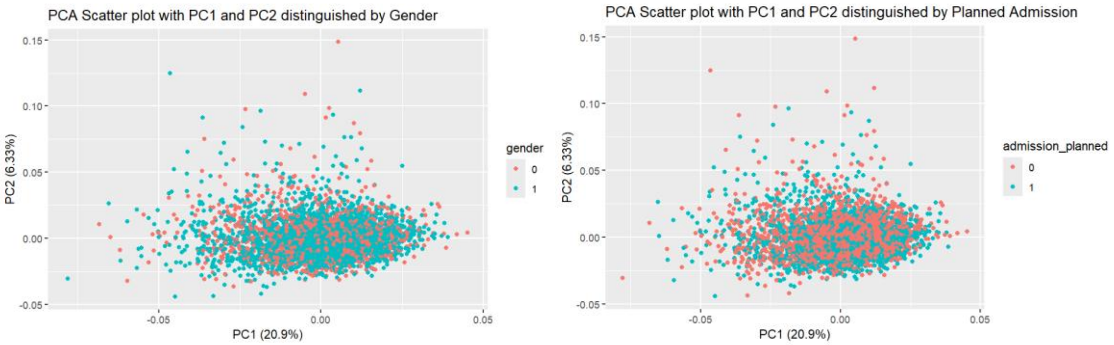


Figure 1. PCA Scatter plot with PC1 and PC2 distinguished by Gender (left) and planned admission (right).

To enhance the algorithm performance and reduce the risk of overfitting for our model, we sought to fit the model for our training data using the 'best features' through the backward selection approach. Therefore, we incorporated all 143 features to learn the model for the training data. This approach yielded a perfect accuracy of 100% and a kappa of 1.00 on the training dataset (Table 3), indicating a perfect agreement between prediction and reality. Subsequently, because RF allows us to determine feature importance (see R code), we were able to select the most important features to fit the final model. Among the most important features, sedative use, type of coma, and morphine dose were ranked highest in order of importance. Even though this approach slightly reduced the accuracy and the kappa statistic of the final model to 99.99% and 99.87% respectively (Table 3), it yielded an improved model with more generalizable potential in our estimation. To validate this, the model was applied to the test data, and an accuracy of 85% and a kappa of 0.54 was observed (Table 3), signifying a moderate agreement between prediction and reality. These values were slightly higher than those of the model with all features on test data. We observed a high-class error for the minority classes (Table 1 and 2).

Table 1. Confusion matrix of the RF model with 20 features on the train data.

Delirium Classification	No Delirium	Hypoactive Delirium	Mixed Delirium	Hyperactive Delirium	Class error
No Delirium	3667	112	15	2	0.03398314
Hypoactive Delirium	114	284	49	8	0.37382637
Mixed Delirium	70	185	28	5	0.90277778
Hyperactive Delirium	37	124	16	4	0.97790055

Table 2. Confusion matrix of the final RF model with 20 features on the test data.

Delirium Classification	No Delirium	Hypoactive Delirium	Mixed Delirium	Hyperactive Delirium	Class error
No Delirium	3667	115	12	2	0.03398314
Hypoactive Delirium	109	285	54	7	0.37362637
Mixed Delirium	71	186	27	4	0.90625000
Hyperactive Delirium	39	124	14	4	0.97790055

Table 3. The general model parameters of the Random Forest models on train and test data.

Delirium Classification	**Train data**		**Test data**	
	Accuracy	*Kappa*	*Accuracy*	*Kappa*
5-feature Model	0.9307	0.7602	0.7976	0.0316
All-feature Model	1.0000	1.0000	0.8439	0.4842
20-feature (final) Model	0.9996	0.9987	0.8513	0.5381
Weighted Random Model	1.0000	1.0000	0.8050	0.0043

Table 4. The sensitivity and specificity for all delirium classes of the Random Forest model on the test data.

Delirium Classification	**Sensitivity / Specificity**	**Interpretation**
No Delirium	0.9688 / 0.7758	Very high / High
Hypoactive Delirium	0.7228 / 0.9015	Moderate / High
Mixed Delirium	0.0857 / 0.9878	Very low / Very high
Hyperactive Delirium	0.0137 / 0.9959	Very low / Very high

Implications

As discussed in the motivation section, our prediction for high accuracy for RF was observed. However, there was a bias in predictions toward the majority class in our test data. This indicates that predicting delirium for a randomly selected patient—whether they exhibit delirium or not, and particularly the delirium subtype—poses challenges and may not be reliable. We sought to alleviate this problem through the application of cost-sensitive learning, wherein we assigned class weights, giving higher weights to the minority classes, as demonstrated in the R code. This, however, did not improve the model performance (Table 3), but it is a future direction that could be explored. Future work may focus on using the balanced random forest approach to improve predictive accuracy and recall, especially since sensitivity in minority groups was found to be relatively low (Table 4). Lowering the decision threshold and tuning hyperparameters could additionally enhance recall. Finally, applying Ridge Regularized Regression may further boost model performance by mitigating overfitting and multicollinearity.

References

1. Chen W, Yang K, Yu Z, Shi Y, Chen CLP. A survey on imbalanced learning: latest research, applications and future directions. Artificial Intelligence Review. 2024;57(6):137.
2. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
3. Chen C, Breiman L. Using Random Forest to Learn Imbalanced Data. University of California, Berkeley. 2004.
4. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Medical Informatics and Decision Making. 2011;11(1):51.