Supplementary Digital Content 1. A brief description of each machine learning algorithm and the parameters that can be tuned for each algorithm using the R statistical programming language.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Description | Parameters | Parameter description |
| Naïve Bayes | A probabilistic classifier based on Bayes’ theorem and conditional probabilities | Laplace correction | A technique used to smooth categorical data and prevent the high influence of zero probabilities |
|  |  |  |
|  | Distribution type | Specifies whether a kernel density estimator is used or not |
|  |  |  |
|  | Bandwidth adjustment | Specifies the bandwidth size. Only relevant when a kernel density estimator is used |
|  |  |  |
| Logistic regression | Used to develop a regression equation from the log odds | No tuning parameters |  |
|  |  |  |  |
| Random forest | An ensemble learning method that operates by constructing multiple decision trees | Number of randomly selected predictors | Specifies the number of variables available to be used to split the data at each tree node |
|  |  |  |  |
| Support vector machine | Used to construct a non-linear, multidimensional separating hyperplane for classifying observations | Sigma | Specifies the flexibility of the decision boundary |
|  |  |  |
|  | Cost | Specifies the cost of incorrect classification |
|  |  |  |
|  | Weight | Allocates a weighting to a preferential class |
|  |  |  |
| Neural network | A network of connected artificial neurons which transmit information and learn from correct or incorrect classifications | Number of hidden units | Number of hidden units used to apply functions to the input data |
|  |  |  |
|  | Weight decay | Specifies the cost of allocating large weights to input units |
|  |  |  |