

CASE STUDY: Predict Disease Deaths with ML and R

We used the outbreaks package that contains datasets on epidemics outbreak of Influenza data as an example to test whether we can use Machine Learning algorithms for predicting disease outcome.

To do so, BigDataOi selected and extracted features from the raw data, including age, days between onset and outcome, gender, whether the patients were hospitalized, etc. Missing values were imputed, and different model algorithms were used to predict outcome (death or recovery). The prediction accuracy, sensitivity, and specificity. The thus prepared dataset was divided into training and testing subsets. The test subset contained all cases with an unknown outcome. Before BigdataOi applied the models to the test data, BigdataOi further split the training data into validation subsets.

The tested modeling algorithms were similarly successful at predicting the outcomes of the validation data. To decide on final classifications, BigdataOi compared predictions from all models and defined the outcome "Death" or "Recovery" as a function of all models, whereas classifications with a low prediction probability were flagged as "uncertain". Accounting for this uncertainty led to a 100% correct classification of the validation test set.

The training cases with unknown outcome were then classified based on the same algorithms. From 57 unknown cases, 14 were classified as "Recovery", 10 as "Death" and 33 as uncertain.

