Comparing the results of the three learning algorithms

Algorithm	Accuracy(%)	Precision(%)	Recall(%)	Model Construction Time(s)
Decision Tree	84%	73%	58%	0.2
Gradient Boosting	85%	77%	58%	10.3
Random Forest	84%	73%	58%	6.0

Summary

The overall accuracy in identifying whether a case was related to an outbreak or not was 84-85%% for all three models. The models all performed similarly, all of them struggled to identify outbreak related cases. The only important feature was age. There were many other “weak learners”.

We would rank the models as:

1. Gradient Boosting
2. Decision Tree
3. Random Forest

In terms of quality Gradient Boosting provides slightly higher accuracy and precision with recall being similar to the other two models. Decision Tree and Random Forest perform almost identically. The Decision Tree algorithm however runs much faster which is the motivation behind ranking it above Random Forest. These results can be explained by Gradient Boosting doing well with weak learners relative to Random Forests & Decision Trees.

We learned that outbreaks have a strong relation with age groups rather than any of the mobility or climate features. Our data likely doesn’t provide enough information regarding whether a case is outbreak related or not. We can conclude that mobility & weather related information is not sufficient to predict outbreak related cases. Alternative data sources might have been location and socioeconomic factors that could be more indicative of outbreak related cases.