**Model validation metrics**

For each model as they use a different subset of data—training, and internal and external testing data sets—we calculated the confusion matrix of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), and derived the four metrics to evaluate the model comprehensively:

1. Accuracy = correct classifications / total classifications = (TP + TN)/(TP+TN+FP+FN)

Accuracy involves all four components of the confusion matrix, and usually serves as a coarse measure of overall model quality. In case of imbalanced data, such as ours where there are fewer high salinity points (30 out of 106) than there are for lower salinity (70 out of 106), there is a possibility of getting a relatively high accuracy even if all the high salinity points are incorrectly classified as low (accuracy would still be 70% because TN=70 even when FN=30). We, therefore, also consider the following three metrics in tandem.

1. Precision = correctly classified positives / all classified as positive = TP / (TP+FP)

Precision improves when false positives decrease, that is, high saline data is not classified as low.

1. Recall = correctly classified positives / all actual positives = TP / (TP+FN)

Recall, or sensitivity or true positive rate, improves when false negatives reduce (that is fewer low saline points are classified as high).

1. F1-score = 2 \* (precision \* recall) / (precision + recall)

F1 score is the harmonic mean of precision and recall, and higher score suggests a more balanced model.