Submarine Avoid Mines

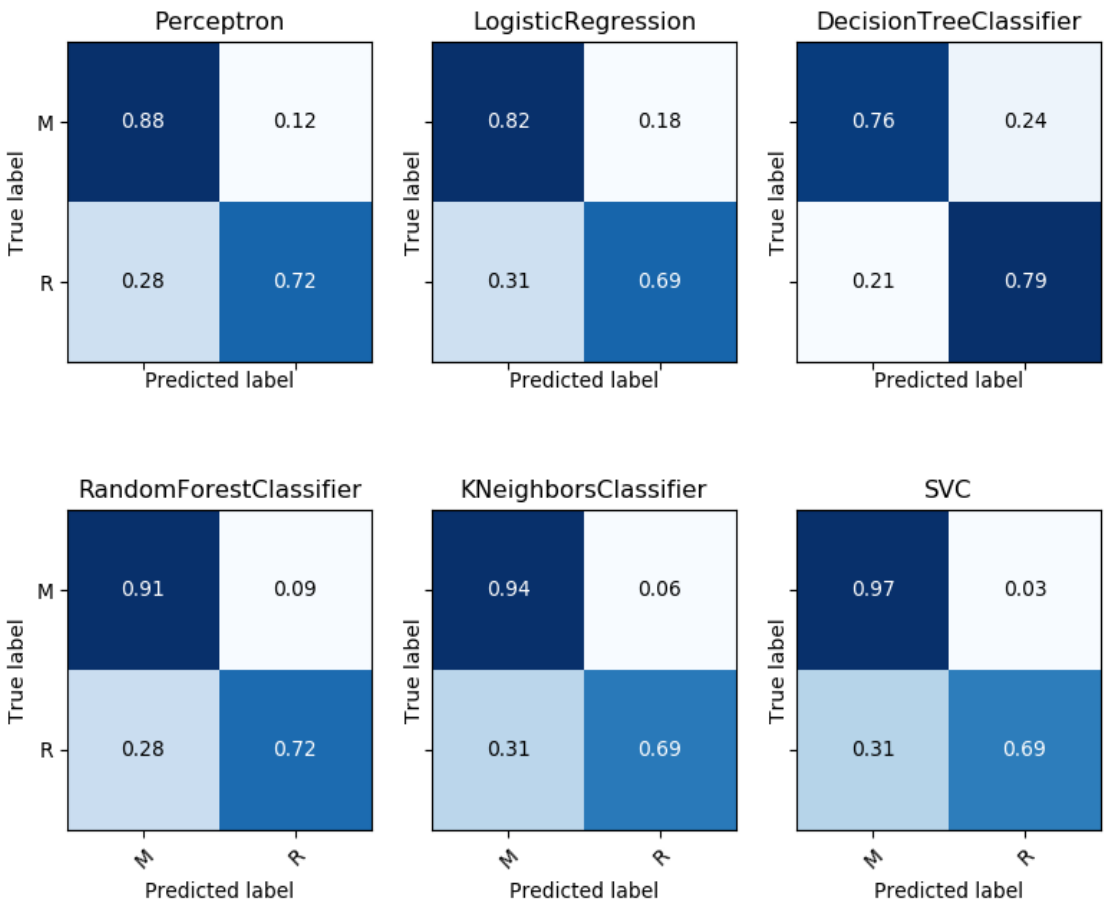
Our current problem in hand is to avoid mines in sea in the path of our submarine. Our mission here is to differentiate mines from rocks in the path. Going close to the mines is fatal as our submarine is old. We have been given a dataset of 208 X 62 dimensions - with the last column indicating whether it's a mine or a rock. Here, since our dataset is labelled, we can use supervised machine learning to solve our problem. Therefore in our analysis, we explored several machine learning techniques to analyze data and make predictions. One thing that is critical in our study is to avoid falsely classifying mines as rocks as this is fatal. Wrongly predicting mines as rocks can cause our submarine to hit them. However, wrongly classifying rocks as mines is permissible as it's not directly fatal to us. Therefore, in our analysis we have plotted "confusion matrices" to check these conditions. We have also generated a consolidated report for different classifiers in our study so that we can compare them easily amongst each other.

Since the number of features is high, we have used Principal component analysis is based to provide dimensionality reduction with minimum loss of information. The wide range of features here might be highly correlated between each other, and models can easily over fit. PCA helps us reduce this. After trying lot of values, we found that using PCA with 30 dimensions performing significantly well. Therefore, we went ahead with this.

Our Result:

Classifier	Result
Perceptron	Misclassified:12 Accuracy:0.8095238095238095 Combined Accuracy:0.8461538461538461
Logistic Regression	Misclassified:15 Accuracy:0.7619047619047619 Combined Accuracy:0.8557692307692307
Decision Tree Classifier	Misclassified:14 Accuracy:0.7777777777777778 Combined Accuracy:0.9326923076923077
SVM	Misclassified:10 Accuracy:0.8412698412698413 Combined Accuracy:0.9519230769230769
K-Neighbors Classifier	Misclassified:11 Accuracy:0.8253968253968254 Combined Accuracy:0.9182692307692307
Random Forest Classifier	Misclassified:11 Accuracy:0.8253968253968254 Combined Accuracy:0.9471153846153846

Confusion Matrices



Here, we can see that Perceptron being a very basic classifier performs very mediocre with lot of mines being classified as rocks.

Logistic Regression however, though simple performs considerably well in a space as the classifier here is binary in nature being a non-linear classifier with a linear combination of parameters and nonlinear function.

Decision tree too performs very mediocre. Hence we rarely use a single decision tree and always go with multiple decision tree classifiers such as Random Forest Classifier.

We can see that Random Forest Classifier with 100 tree classifiers performs much better. This is because Random forest uses averaging to improve the predictive accuracy and control over-fitting.

In K-nearest neighbors' classifier, we try to predict based on the values of the nearest neighbors to a point. Here, we got best results while using 'minkowski' metric for distance calculation and used 3 neighbors for classifying the points.

Amongst all the classifier Support Vector Machine was found to be the best. SVM tries to find a hyper plane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. We tried various kernels, gamma values and decision function shapes. And we were able to finalize on the best SVM classifier while using the 'ovo' (one-vs-one) decision function shape. Another convincing reason to use SVM is the very less Mines being classified as Rocks (0.03%) therefore increasing safety of our journey considerably. Hence we would like to finalize our solution with an SVM based system.