Setting up packages and data

Load Required packages.

```r
library(ROCR) # ROC analysis library(caret) # Statistical Learning library(MASS) library(RCIA)
```

Load data and split training data to create a test set.

```r
data("glassworm", package = "mlbench")
set.seed(1)
trainSet <- glassworm[1:500,]
index <- createDataPartition(glassworm$Class, p = 0.7, list = FALSE)
trainSet <- glassworm[index,]
testSet <- glassworm[-index,]
```

**Feature Selection**

We will use a simple feature selection method: **RFE**. Recursive feature elimination is used to find the top performing features with a `randomForest` model. Cross-validation will be used to test performance.

```r
set.seed(1)
control <- rfeControl(functions = rfFuncs, method = "cv", number = 10)
trainSet$x <- scale(trainSet$x)

rf_model <- randomForest(Class ~ ., data = trainSet, control = control)
summary(rf_model)
```

<table>
<thead>
<tr>
<th>Variables</th>
<th>ROC</th>
<th>Sens</th>
<th>Spec</th>
<th>ROCSD</th>
<th>SensSD</th>
<th>SpecSD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.818607</td>
<td>0.763033</td>
<td>0.730892</td>
<td>0.011246</td>
<td>0.153209</td>
<td>0.177104</td>
</tr>
<tr>
<td>5</td>
<td>0.8091603</td>
<td>0.8268774</td>
<td>0.8857143</td>
<td>0.0082360</td>
<td>0.1421846</td>
<td>0.1675422</td>
</tr>
<tr>
<td>10</td>
<td>0.9249698</td>
<td>0.871429</td>
<td>0.860476</td>
<td>0.0182927</td>
<td>0.1340977</td>
<td>0.1425261</td>
</tr>
<tr>
<td>15</td>
<td>0.9127279</td>
<td>0.8648571</td>
<td>0.8574016</td>
<td>0.0359630</td>
<td>0.1510963</td>
<td>0.1605290</td>
</tr>
<tr>
<td>20</td>
<td>0.9321329</td>
<td>0.871429</td>
<td>0.8748286</td>
<td>0.0383062</td>
<td>0.1566640</td>
<td>0.1420513</td>
</tr>
<tr>
<td>25</td>
<td>0.9309524</td>
<td>0.871429</td>
<td>0.860476</td>
<td>0.0279508</td>
<td>0.1510963</td>
<td>0.1425261</td>
</tr>
<tr>
<td>30</td>
<td>0.9319728</td>
<td>0.871429</td>
<td>0.8574016</td>
<td>0.0383092</td>
<td>0.1510963</td>
<td>0.1312767</td>
</tr>
<tr>
<td>35</td>
<td>0.9472109</td>
<td>0.8857143</td>
<td>0.8574016</td>
<td>0.0896511</td>
<td>0.130427</td>
<td>0.1505567</td>
</tr>
<tr>
<td>40</td>
<td>0.9363046</td>
<td>0.871429</td>
<td>0.8648571</td>
<td>0.0844812</td>
<td>0.1510963</td>
<td>0.1425261</td>
</tr>
<tr>
<td>45</td>
<td>0.9276912</td>
<td>0.871429</td>
<td>0.890243</td>
<td>0.0924731</td>
<td>0.1566640</td>
<td>0.1176014</td>
</tr>
<tr>
<td>50</td>
<td>0.9323313</td>
<td>0.871429</td>
<td>0.890243</td>
<td>0.0871047</td>
<td>0.1566640</td>
<td>0.1176014</td>
</tr>
<tr>
<td>55</td>
<td>0.9102878</td>
<td>0.871429</td>
<td>0.890243</td>
<td>0.0889900</td>
<td>0.1549093</td>
<td>0.1088866</td>
</tr>
<tr>
<td>60</td>
<td>0.9316163</td>
<td>0.871429</td>
<td>0.890243</td>
<td>0.0965520</td>
<td>0.1340977</td>
<td>0.1176014</td>
</tr>
</tbody>
</table>

Plot results of cross-validation with different numbers of features used.

```r
plot(results, type="f", xlab="f")
```

Test results

You can then compare the model built with all the features or with the selected from feature selection and compare performance on test data.

```r
set.seed(1)
control <- rfeControl(functions = rfFuncs, method = "cv", number = 10)
trainSet$x <- scale(trainSet$x)

rf_model <- randomForest(Class ~ ., data = trainSet, control = control)
 predictions <- predict(rf_model, testSet, type = "prob")
```

Summary of performance with different number of features selected.

```r
#feature selected data performed better on training and provided equivalent performance on a test set
#compare performance on test data.

library(gmodels)

roc_table <- data.frame(Training = Training, Training Feature Selected = Training Feature Selected, Testing = Testing, Testing Feature Selected = Testing Feature Selected)

roc_table
```

```r
Data | AUC
--- | ---
Training | 0.936
Training Feature Selected | 0.955
Testing | 0.883
Testing Feature Selected | 0.883
```