Introduction

The data measures four criteria (dimensions) of 150 human skulls found in Egypt, and assigns the observations to one of five time periods (epochs) based on the measurements. Interestingly, the observations are divided equally between the five epochs (so, 30 observations each). Initially, the maxima and minima of each of the criteria in each epoch were examined (**table 1**). What is noticeable is the relative closeness of each of the ranges in each of the criteria, making it difficult to identify clear values that differentiate each of the epochs. A cursory examination would suggest that the 4000BC epoch is marked by lower mb and higher bl values than the others, whilst 3300BC shows wide bounds for bh and narrow ones for nh. Similarly, the 150AD epoch shows slightly higher maximum values for mb and slightly lower minima for nh. All of this points to possible future challenges in fitting classifiers to the dataset.

Epoch	mb	bh	bl	Nh
4000BC	119-141	121-143	89-114	47-56
3300BC	124-145	124-145	90-107	45-56
1850BC	126-140	125-142	87-106	45-60
200BC	129-144	120-141	86-107	46-60
150AD	126-147	120-136	81-103	44-58

Table 1 – Maximum and minimum values of each of the classification critieria

Initial Classification

This was underlined by fitting a classification tree to the data. The predicted and actual values match about 50% of the time when the same data was used for training and test (**figure 1**). Moreover, the tree in question was somewhat complex and fragmented resulting in 6 levels and a total of 12 bottom level nodes, with observation numbers ranging from 8 to 28 skulls – with the majority of nodes having 8 – 12 observations. When a 75%/25% split was made between training and test data, the accuracy of the model fell to around 20% (**figure 2**). When repeated with a simulation of 100 values, the results remain roughly the same with model accuracy on the training data falling below 50% (**figure 3**).

```
> # Look at table (rows=truth, cols=prediction)
> tab <- table(skulls$epoch,pred)
> tab
         pred
          c4000BC c3300BC c1850BC c200BC cAD150
  c4000BC      15       7       4      2      2
  c3300BC       5      21       0      1      3
  c1850BC       5       6      16      1      2
  c200BC        2       4       6     11      7
  cAD150        1       3       8      4     14
>
> # Work out the accuracy
> sum(diag(tab))/sum(tab)
[1] 0.5133333
>
```

Figure 1 – Initial Classification (full dataset)

```
>
> # Look at table for the test data only (rows=truth
> tab <- table(skulls$epoch[indtest],pred[indtest])
> tab

         c4000BC c3300BC c1850BC c200BC cAD150
  c4000BC       2       3       1      1      2
  c3300BC       2       1       2      1      1
  c1850BC       1       3       1      0      6
  c200BC        0       1       2      1      4
  cAD150        0       1       1      0      3
>
> # Work out the accuracy
> sum(diag(tab))/sum(tab)
[1] 0.2
>
```

Figure 2 – Initial Classification (Test data only)

```
> #Check out the error ra
> colnames(res)<-c("test"
> apply(res,2,summary)
            test  train
Min.     0.05263 0.4018
1st Qu.  0.18420 0.4643
Median   0.21050 0.4732
Mean     0.21290 0.4780
3rd Qu.  0.23680 0.5000
Max.     0.39470 0.5536
```

Figure 3 – Initial Classification (simulation of 100 values)

Finally, use of a 10 fold cross validation approach yielded an accuracy of 23%, concurring with the general trend (**figure 4**).

Further measures of classifier accuracy

When other measures of classifier accuracy were used, two trends became apparent. Firstly, the accuracy rate based on test data remained at around the 20-30% mark, only marginally higher than before. Secondly, the accuracy rate based on training data only increased markedly – the extent depending on the measure used. An example of this was when bagging was used, where the accuracy for the test data reached close to 29%, and 60% when assessed with the training dataset (**figure 5**). When Random Forest was used, the results were barely different for the test data, although they reached a 100% accuracy rate using the training dataset (**figure 6**). Whilst this might appear promising, it is the low accuracy on the test data (<30%) that remains worrying.

```
            test
Min.     0.06667
1st Qu.  0.15000
Median   0.26670
Mean     0.23330
3rd Qu.  0.31670
Max.     0.40000
> res
             test
 [1,] 0.26666667
 [2,] 0.26666667
 [3,] 0.06666667
 [4,] 0.26666667
 [5,] 0.40000000
 [6,] 0.20000000
 [7,] 0.33333333
 [8,] 0.13333333
 [9,] 0.06666667
[10,] 0.33333333
> table(folds)
folds
 1  2  3  4  5  6  7  8  9 10
15 15 15 15 15 15 15 15 15 15
```

Figure 4 – Cross Validation (k=10)

```
> # Test data
> table(skulls$epoch[indtest],pred.b[indtest])

         c1850BC c200BC c3300BC c4000BC cAD150
  c4000BC       4      2       1       1      0
  c3300BC       0      0       3       3      0
  c1850BC       3      1       2       0      0
  c200BC        3      1       1       0      4
  cAD150        1      3       0       2      3
> sum(skulls$epoch[indtest]==pred.b[indtest])/length(indtest)
[1] 0.2894737
>
> # Training data
>
> table(skulls$epoch[indtrain],pred.b[indtrain])

         c1850BC c200BC c3300BC c4000BC cAD150
  c4000BC       3      1       6      12      0
  c3300BC       1      2      19       2      0
  c1850BC      16      1       3       3      1
  c200BC        1     11       6       0      3
  cAD150        6      0       4       1     10
> sum(skulls$epoch[indtrain]==pred.b[indtrain])/length(indtrain)
[1] 0.6071429
>
```

Figure 5 – Bagging

```
> # Test data
> table(skulls$epoch[indtest],pred.rf[indtest])

         c4000BC c3300BC c1850BC c200BC cAD150
  c4000BC       1       1       5      1      0
  c3300BC       4       2       0      0      0
  c1850BC       1       3       1      1      0
  c200BC        0       0       3      3      3
  cAD150        2       1       1      2      3
> sum(skulls$epoch[indtest]=pred.rf[indtest])/le
```

```
> # Training data
> table(skulls$epoch[indtrain],pred.rf[indtrain])

         c4000BC c3300BC c1850BC c200BC cAD150
  c4000BC      22       0       0      0      0
  c3300BC       0      24       0      0      0
  c1850BC       0       0      24      0      0
  c200BC        0       0       0     21      0
  cAD150        0       0       0      0     21
```

Figure 6 – Random Forest (Test and Training data)

Conclusion

Regardless of whichever classifier is used the predicted accuracy of the model, on test rather than training data, does not go above 20-30%. Moreover, the approaches involving classification trees struggle to place the observations into a manageable number of nodes. The persistent low accuracy of the predictive models is likely to derive from the nature of the data in question. This includes factors such as:

1. **Discrimination** – The existing five groups are not well defined. The variables are all within the same narrow bands regardless of whichever epoch they are assigned to. It is difficult to tell what makes the measurements for one epoch different to those of another. Because of this, the software has difficulty in discriminating between them.
2. **Additional Variables** – There are only four variables that the model can use to construct the model. Extra variables, if sufficiently distinct, would provide the model with a more robust, and possibly simpler way of classifying the variables into coherent and meaningful groups.