DNA DATA SET

Introduction

Bacteriophages, or just phages, are viruses that are the natural predators of bacteria and argued to be the most abundant entities on earth. They have been found to contribute to the evolution of bacterial cells within the human body by acquiring and spreading DNA. Despite their considerable presence in the human body, it is not clear how they impact human health.

In this report we aim to create a model for distinguishing between human and phage DNA sequences. To achieve this aim we will create decision trees and random forests which will operate using DNA base information from our dataset of 600 (300 human and 300 phage) DNA sequences. We hypothesise that human and phage DNA follow different sequence patterns and that by breaking up combinations into individual base numbers we can identify phage DNA from human DNA.

Data and Methods

The Data

The data set is comprised of 300 human and 300 phage DNA sequences which are labelled as 'pos' and 'neg' respectively in the first column of the data set. The next 100 columns contain the four bases which make up the DNA sequences (represented by rows). This information was collated into second data frame through feature extraction to help us examine whether the number of bases would be better at distinguishing between human and phage DNA sequences than using sequences as a whole.

The variables are summarised in the table below.

Variable	Description
numAs	number of A bases
numCs	number of C bases
numGs	number of G bases
numTs	number of T bases
numATs	number of AT bases
numCGs	number of CG bases
dna.V1	human phage DNA sequence (pos/neg)

Methods

Decision Trees – Decision trees are a supervised learning method that can be used for solving regression and classification problems. They break down a data set (categorical and numerical data) into many subsets and build up a 'decision tree' with decision nodes and lead nodes. A decision node has two or more branches, and a leaf node represents a classification. This 'tree' is a training model that ultimately predicts the class or value of a target variable by learning simple decision rules inferred from prior data. This simplicity and easy interpretability make them a good analysis method for our large dataset and will be used in our analysis to distinguish between human and phage DNA sequences.

Pruning – Once decision trees are built, they can become unwieldy and require pruning to remove unnecessary decision node splits which can lead to overfitting the dataset. We prune the model to improve its performance on new data, or in our case, the test set. The amount of pruning we carry out depends on the misclassification rate – we want the smallest misclassification rate relative to the tree size.

Random Forest, Bagging and Boosting – We use ensemble methods which combine multiple decision trees to get more accurate and stable predictions as single decision trees are known to suffer from bias and variance. In our analysis we will be using these ensemble methods to train our model.

Bagging aims to reduce the variance of a decision tree by creating multiple subsets of data from the training sample chosen randomly with replacement. This creates a collection of different models whose prediction are averaged out. Random forest works similarly but extends bagging by also taking random selections of features as well as the random subsets of data. Boosting on the other hand is a sequential method where each model in the sequence is fitted with a greater focus on error so that trees with lower misclassification rates become more important in the final tree construction.

Results and Discussion

We randomly allocated 70% and 30% of the dataset to the training set and test set, respectively and constructed the decision tree in Figure 1 on the training set. This initial decision tree is unwieldly and requires pruning but is used as a baseline model for classifying the data in our training set and test set so that we can begin to differentiate between human and phage DNA. The performance of this initial model on the training data and test data is shown in table 1 and 2 respectively.

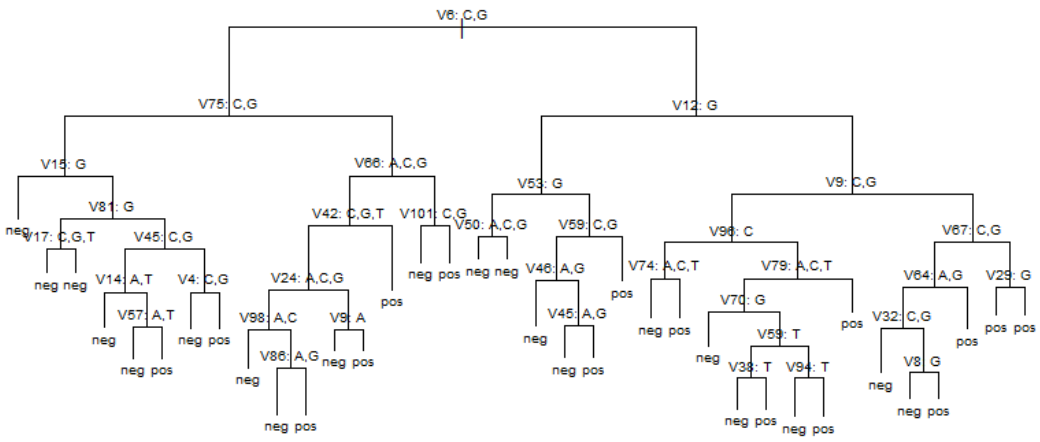

Fig 1: Decision tree created from DNA training set data

		Real	
		Positive	Negative
Predicted	Positive	197	24
	Negative	13	186

Table 1: Confusion matrix obtained from decision tree on training data

		Real	
		Positive	Negative
Predicted	Positive	43	35
	Negative	47	55

Table 2: Confusion matrix obtained from decision tree on test data

We observe that the accuracy of the model on the training data is 91.2% and the accuracy on the test data is 54.4%. Its clear that the model does not generalise well to new data and appears to overfit the data. By looking at the cross-validation plot, we see that the misclassification rate can be reduced with the optimal number of leaf nodes being thirteen as shown in Figure 2.

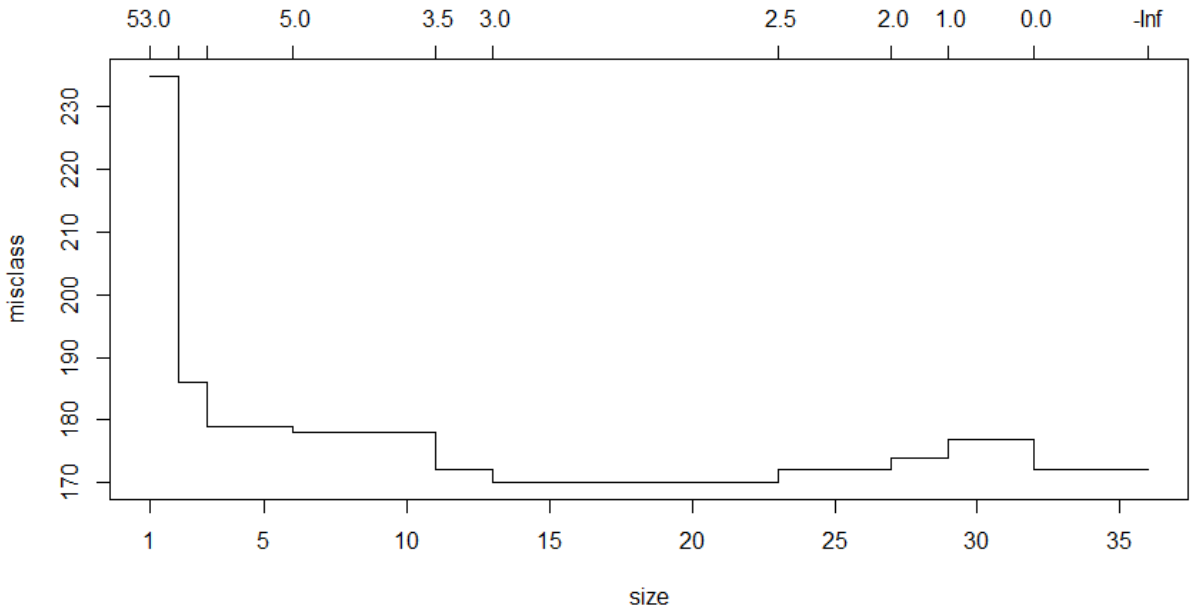

Fig 2: Cross-validation plot showing the misclassification rate against number of leaf nodes

Our revised decision tree with thirteen leaf nodes is shown in Figure 3 and the performance of this revised model on the training set and test set is seen in Table 3 and Table 4, respectively.

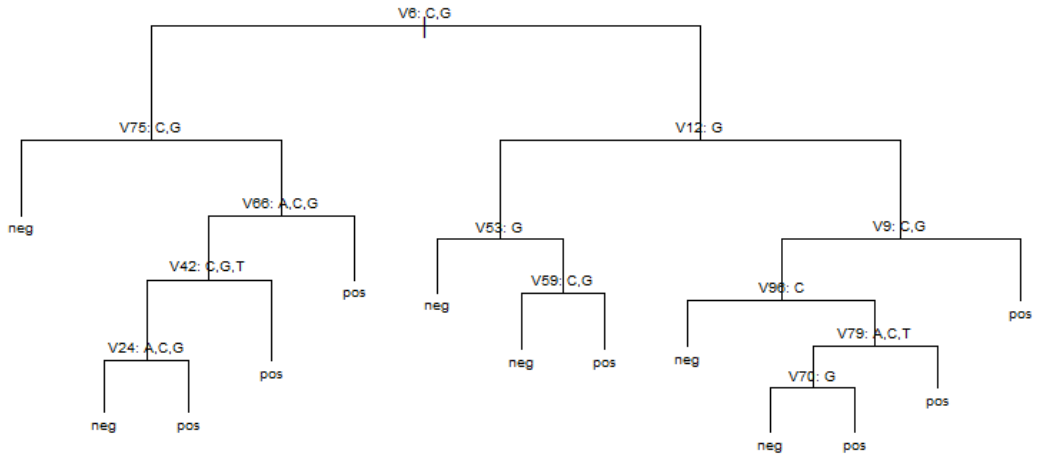

Fig 3: Cross-validation plot showing the misclassification rate against number of leaf nodes

		Real	
		Positive	Negative
Predicted	Positive	176	47
	Negative	34	163

Table 3: Confusion matrix obtained from pruned decision tree on training data

		Real	
		Positive	Negative
Predicted	Positive	51	35
	Negative	39	55

Table 4: Confusion matrix obtained from pruned decision tree on test data

The accuracy of the revised model on the training data is now 80.7% and 58.9% on the test data. This is a slight improvement from the unpruned model, but at an accuracy rate of 58.9% the model is still far from useful.

Our random forest showed a significant improvement in accuracy on the test data at 71.1% shown in Table 5.

		Real	
		Positive	Negative
Predicted	Positive	69	31
	Negative	21	59

Table 5: Confusion matrix obtained from random forest on test data

We repeat the prior steps with the extracted features data set and compare the results.

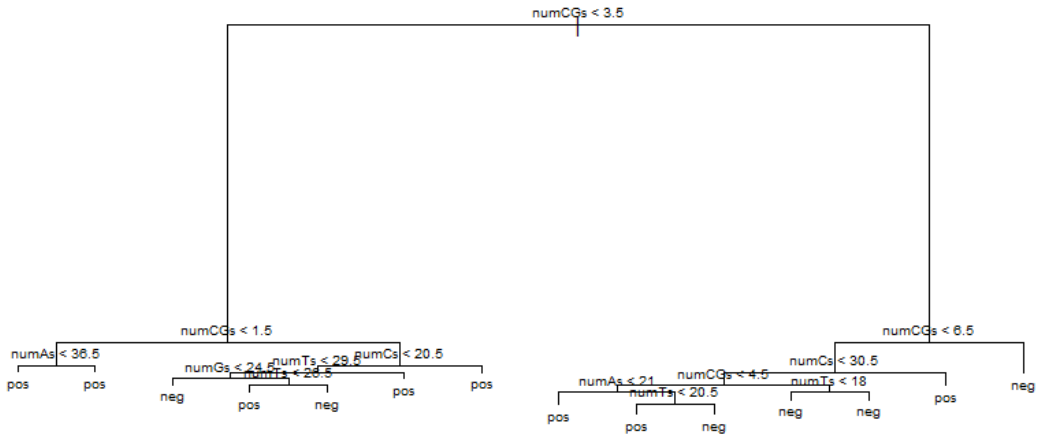

Fig 4: Decision tree created from extracted feature data set

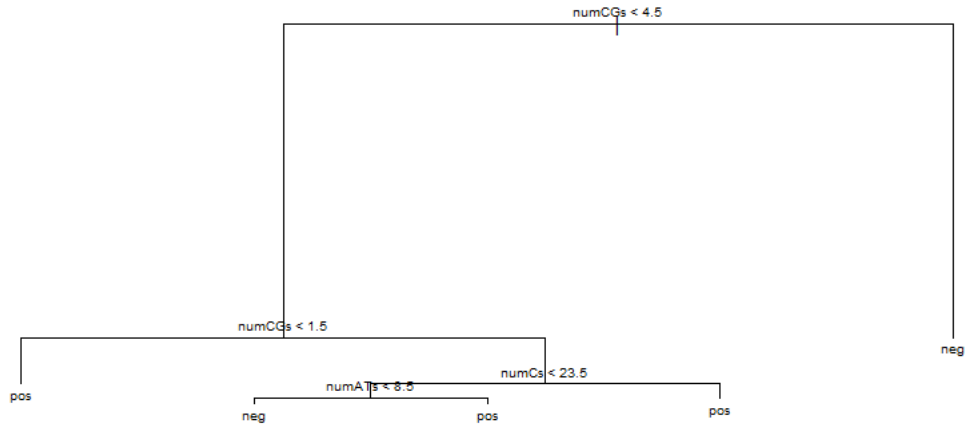

Fig 5: Pruned Decision tree created from extracted feature data set

We observe a great improvement in accuracy using our extracted feature data set compared to our original data set. The performance of our decision tree using the extracted feature data set is 87.8% accuracy on the test set. This figure was improved further to 88.3% accuracy after pruning to 5 leaf nodes. Our random forest on this extracted data set came back with 92.8% accuracy.

Although boosting our extracted feature data set on both the training and test set does not improve our accuracy further, we were able see which variables are driving the performance in Figure 6. We see that 'numCGs' appears to be the most significant DNA base in being able to distinguish between human and phage DNA.

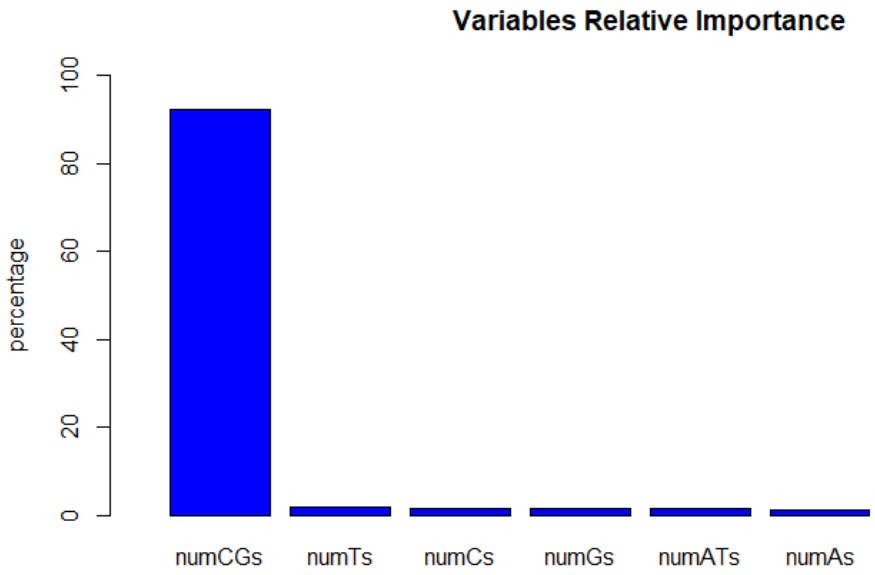

Fig 6: Bar chart showing how much each variable contributes to boosted random forest

Conclusions

We see that the random forest classification on the extracted feature data set performs best with regards to being able to accurately distinguish between human and phage DNA sequences and the feature 'CG' was identified as being the most important variable contributing to the models success. This is because the feature appears to be more likely to belong to the phage DNA sequence group than the human DNA sequence, hence it is crucial in our classification. We observe that the extracted feature data set performed much better than the original dataset despite using the same methods,

The improvement in performance compared to using the original data shows that using number of bases performs better in being able to distinguish between human and phage DNA sequences. This is because both groups of DNA sequences share the same base types, the difference lies in the number of occurrences of each base in a sequence.