Marinus Burger, UU# F132726 Joseph Eremondi, UU# 4229924

Data Set Description: We found the Spambase Dataset as a part of the UCI Machine Learning Repository, a repository focused on classifying emails as spam or non-spam. It contains 57 continuous attributes and one nominal class label attribute which classifies the email. For classifying the emails, for example, the ratio of words in the mail that match WORD to the total number of words in the email was calculated. Words such as “credit” and “money” were used to identify spam, whereas “george” and the area code of Hewlett Packard were among the words used to identify legitimate mail. The Dataset contains a misclassification error around 7%. The total number of Instances is 4601, with 1813 instances classified as spam.

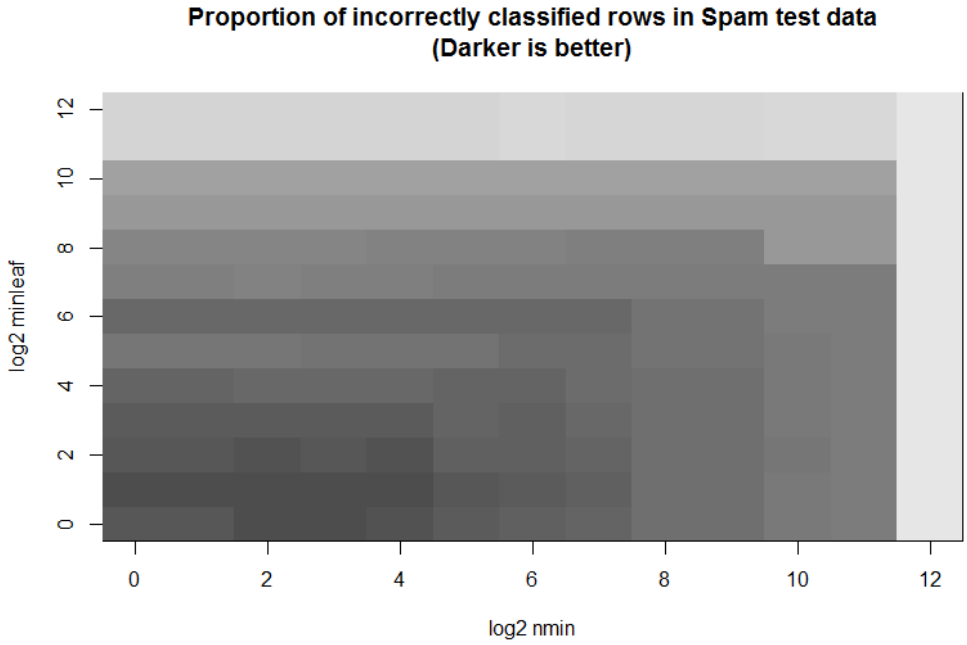
Analysis: We randomly divided the Spam data set into two sets, a training set with 70% of the data, and a testing set with 30% of the data.

We tested the performance of the algorithm on the test data after being trained on the training data, for (nmin, minleaf) = (2^i, 2^j) for every i,j in {0 … 12}. Our tests yielded the following results:

	minleaf												
nmin	1	2	4	8	16	32	64	128	256	512	1024	2048	4096
1	8.26	7.46	8.26	8.84	10.22	12.39	10.58	14.20	15.14	18.77	20.80	34.35	34.35
2	8.26	7.46	8.26	8.84	10.22	12.39	10.58	14.20	15.14	18.77	20.80	34.35	34.35
4	7.61	7.54	8.19	8.91	10.36	12.46	10.65	14.35	15.14	18.77	20.80	34.28	34.28
8	7.68	7.25	8.33	8.91	10.29	12.10	10.43	14.20	15.07	18.77	20.80	34.35	34.35
16	8.04	7.68	8.19	8.77	10.29	11.88	10.29	13.84	14.71	18.77	20.80	34.35	34.35
32	9.20	8.62	9.49	9.93	10.14	11.88	10.29	13.62	14.64	18.77	20.87	34.35	34.35
64	9.42	8.84	9.42	9.71	10.22	10.87	10.29	13.48	14.35	18.70	20.65	35.22	35.22
128	9.78	9.71	9.93	10.36	10.94	11.16	10.36	13.48	14.28	18.55	20.43	34.86	34.86
256	11.59	11.52	11.74	11.67	11.67	11.88	12.25	13.48	14.28	18.55	20.51	34.86	34.86
512	11.59	11.52	11.74	11.67	11.67	11.88	12.25	13.48	14.28	18.55	20.58	34.78	34.78
1024	12.83	12.83	12.75	12.83	12.83	12.83	13.48	13.48	18.55	18.55	20.58	35.22	35.22
2048	13.48	13.48	13.48	13.48	13.48	13.48	13.48	13.48	18.55	18.55	20.58	35.22	35.22
4096	40.07	40.07	40.07	40.07	40.07	40.07	40.07	40.07	40.07	40.07	40.07	40.07	40.07

Incorrectly classified input instances (%)

Additionally, we visualised our data using the image() function of R.



It can be seen that lower values of each parameter tend to provide more accurate predictions than higher values. However, the minimum error rate does not occur with nmin=1 or minleaf=1. Rather, the lowest error rate occurred with nmin=8 and minleaf=2.

Our error rate of 7.25% is comparable to the results obtained by the creators of the data set, who report an error rate of approximately 7%[1].

Having minleaf=2 prevented ever splitting based on a single value. This allowed for impurity reduction, but prevented single outlier values from forming their own leaves. Thus, these noisy values did not incorrectly influence the classification of testing data. Similarly, having nmin=8 prevented the modelling of noise in small data sets.

The classification algorithm still performed very well with minleaf=1 and nmin=1. Our hypothesis is that, because the data is based on word frequencies or real-world emails, there are very few outliers. Internal emails from professionals at Hewlett Packard Labs are unlikely to discuss credit or money, and spam will almost never reference their specific area code. The data set is from 1998, when measures to counteract spam filters were likely undeveloped. So the given data is very specific, and overfitting does not greatly increase error. However, this is subjective intuition, and has not been statistically tested.

The error rates for minleaf < 9 and nmin < 33 are all below 10%. In a real-world scenario where generating small classification trees was a priority, using nmin=32 and minleaf=8 would provide a model which could be stored compactly, but which still had a very high prediction rate.

[1] https://archive.ics.uci.edu/ml/datasets/Spambase