SVM vs Random Forest - A Literature Review and Analysis

CUNY 622 - Assignment 3

Eric Hirsch

10/25/2022

Contents

I. Essay

For this assignment we discuss the advantages and disadvantages of random forest versus SVM, both in the literature and in practice with the previous assignment's data set. As a starting point, two example articles were offered which predicted Covid positive cases using decision tree ensembles and SVM, respectively.

Because of the extreme imbalance between positive and negative cases, the authors of the decision tree ensemble article used decision tree ensembles specifically designed for imbalanced data sets (for example, balanced random forest). Special sampling techniques were also applied to address the imbalance of the data set. The best AUC score (.881) was achieved with RUSbagging.

The SVM article's analysis was slightly more complex, in that it predicted whether a client had no infection, mild infection or serious infection. The model was effective, particularly in predicting severe cases, with an F1 score of .97. The article makes the claim that "SVM works best in predicting COVID-19 cases with maximum accuracy." To support this claim, they performed a comparative study of supervised learning models, including Random Forest. SVM achieved higher F1 scores than the others.

That SVM is the superior model in predicting Covid cases generally is a rather bold statement which can't be justified by one study. There are too many variables – the particular kind and nature of the data collected for the study, the types and tunings of the machine learning algorithms (for example, this study did not consider the " balanced" random forest algorithms that were featured in the first article), and the study design (such as breaking down the classes into no infection, mild or serious).

An examination of articles which compare SVM and random forest in the field of Geography (while I am no longer in this field, it's where I got my education – Berkeley PhD 1996) do not show the superiority of one algorithm over the other. Indeed, there is little consistency either for which algorithm is favored, or for how performance is to be predicted and compared. For example, a study of multispectral images to predict canopy nitrogen weight in corn showed that random forests were only marginally better than SVR (Support Vector Regression) but that the concepts and analysis were easier to interpret with random forests. A study of groundwater mapping showed a higher AUC for random forest than for SVM tested with a linear,

polynomial, sigmoid, and radial kernel. Likewise, a study of sediment transport in the Tigris-Euphrates river also found that random forest predicted better than SVM.

In contrast, a study predicting dominant tree species from Landsat images taken of the Krkonose mountains in the Czech Republic, found that SVM performed better than random forest in that particular study, but acknowledged that random forest consistently provided better results in other studies when spatial resolution was low, while SVM appeared to perform better when there were significantly more features. A study of images by the satellite remote-sensing Sentinel-2A also found that SVM was most accurate (95.17%).

A systematic review conducted in 2020 of 250 peer-reviewed papers on remote-sensing image classification showed that despite some inroads from deep learning, SVM and random forest remained the two most popular image classification techniques, mainly due to lower computational complexity. SVM is seen to be particularly effective where there is high dimensionality and limited training samples, while random forest is easier to use (fewer hyper parameters to tune) and more flexible with more complex classifications. Both tend to be highly accurate, although, some researchers still tout one or the other as the more superior without any strong basis.

As the researchers pointed out, SVM largely depends on the selection of the right kernel function – radial and polynomial tend to be popular in the remote-sensing field. The binary nature of SVM also creates complications for its use in multiclass scenarios, which occur frequently in remote-sensing, although these can be overcome. The ability of SVMs to manage small training sets and operate within high dimensional spaces (which is particularly applicable to remote-sensing image data) make SVMs attractive.

Random Forest, on the other hand, is popular because the decision-making process behind it is clearer and more understandable, it is easily implemented in parallel structure for geo-big data computing, it can handle thousands of input variables, is robust to outliers and noise, and is computationally lighter than other tree ensemble methods.

These recommendations echo those of the machine learning literature. Both algorithms are ... that random forest is better with ... while SVM is better when ...

Beyond these general recommendations, however, the superiority of one algorithm over the other may depend on the idiosyncrasies of the data set you're looking at. It is wise to examine the performance of both before assuming one will perform better than the other. We can illustrate some of this with the database from the previous assignment.

The database contains 466 records of small towns in the Northeast, together with statistics on poverty, industrialization, pollution, crime rate and so on. In the previous assignment, we performed regression analysis on the pupil teacher ratio using random forest. We determined the root mean squared error of predictions on attests that to be .75. An examination of residuals found them to be normally distributed – since the mean is 18.4, this RMSE suggested that the algorithm was effective in distinguishing between high and low people-teacher ratios. For the purposes of comparison with SVM, we created a binary variable containing high PT ratio (the mean and above), and low PT ratio (below the mean.) We ran three support vector machines (linear, radial, and polynomial) as well as a random forest and a decision tree. All of the algorithms were optimized for the parameters which may be tuned.

When all of the variables were included in the data set, random forest performed significantly better than SVM (98% AUC vs. 84% for SVM using a radial kernel). Even a simple decision tree yielded a higher AUC (85%).

We investigated possible reasons that SVM performs so poorly in this data set. It was found that for certain variables (pollution being a prime example), there were anomalous cohorts that were driving the analysis. For example, there was a sizable cohort of schools all at a PT ratio of exactly 21. Further, all of the schools above a pollution index of .65 were in this cohort. There is no indication in the data set documentation for what this cohort means but it may relate to state regulations capping the number of pupils in the class.

Although pollution was barely linearly correlated with PT ratio (1.7), the random forest algorithm relied heavily on it, as anomalies like these lend themselves well to a binary decision point. Indeed, using pollution alone, random forest achieved a 98% AUC on predictions of the evaluation set. The SVM algorithm

(AUC=82%), which only looks at the support vectors, did not appear to benefit from this information. The evidence for this was that when the cohort was removed, the AUC for random forest dropped significantly while the AUC for SVM remained the same. In addition, when the algorithms were run on a data set with minimal columns which were not affected by the "anomalous cohort" phenomenon, radial SVM outperformed all of the other algorithms and polynomial SVM was second.

We also tested the assertion that SVM performs better in high dimensional, low sample data by taking only every fifth record in the data set, reducing it to 94 rows with 12 columns. In this case, SVM does in fact catch up to Random Forest, producing almost the same results.
This analysis shows that it's important to test both algorithms in the data as the reason for the superiority of one performance over the other may not be readily apparent. It can, however, be worthwhile to investigate further what it is about the data shape that favors one over the other.

II.

In this part of the assignment we compare SVM to the Random Forest analysis we performed last week.

####. A. Description

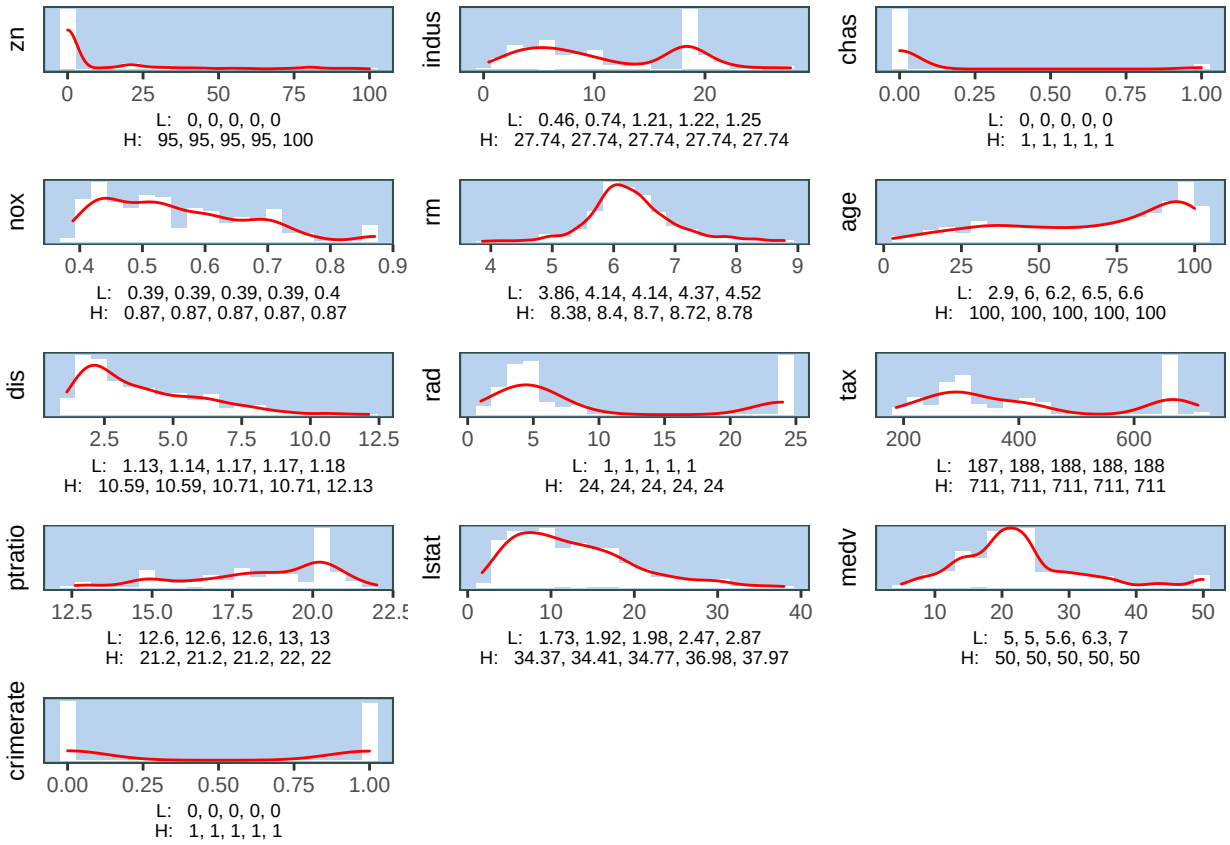
The data set consists of 466 observations of data realted to small towns in the North East. There are 11 numeric variables and two binary variables. There are no missing values. The variables include the level of industrialization, average tax rates, pollution levels, and so on. This data set is often used to predict crime rates, but we won't use it for that purpose.

These are the variables:

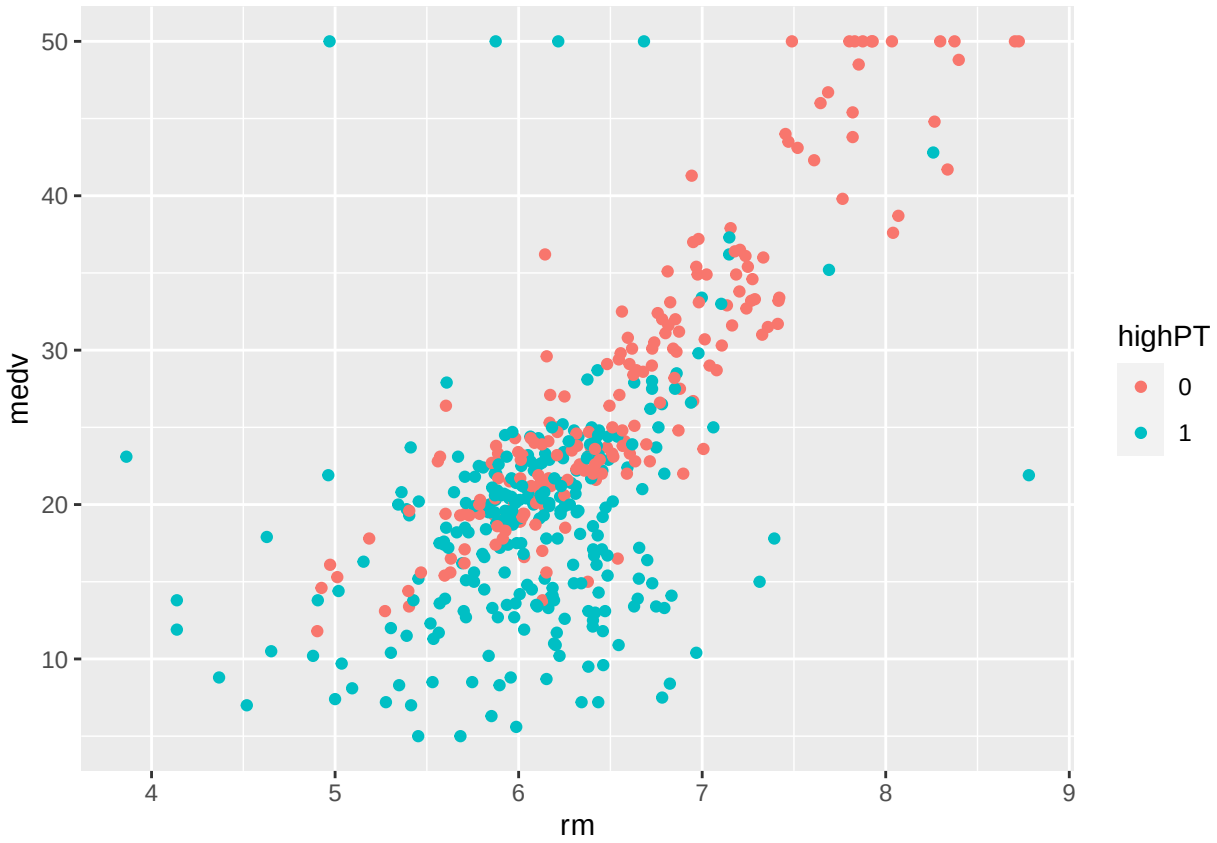
- zn: proportion of residential land zoned for large lots (over 25000 square feet)
- indus: proportion of non-retail business acres per suburb
- chas: a dummy var. for whether the suburb borders the Charles River (1) or not (0)
- nox: nitrogen oxides concentration (parts per 10 million)
- rm: average number of rooms per dwelling
- age: proportion of owner-occupied units built prior to 1940
- dis: weighted mean of distances to five Boston employment centers
- rad: index of accessibility to radial highways
- tax: full-value property-tax rate per $10,000
- ptratio: pupil-teacher ratio by town
- lstat: lower status of the population (percent)
- medv: median value of owner-occupied homes in $1000s
- crime: whether the crime rate is above the median crime rate (1) or not (0)

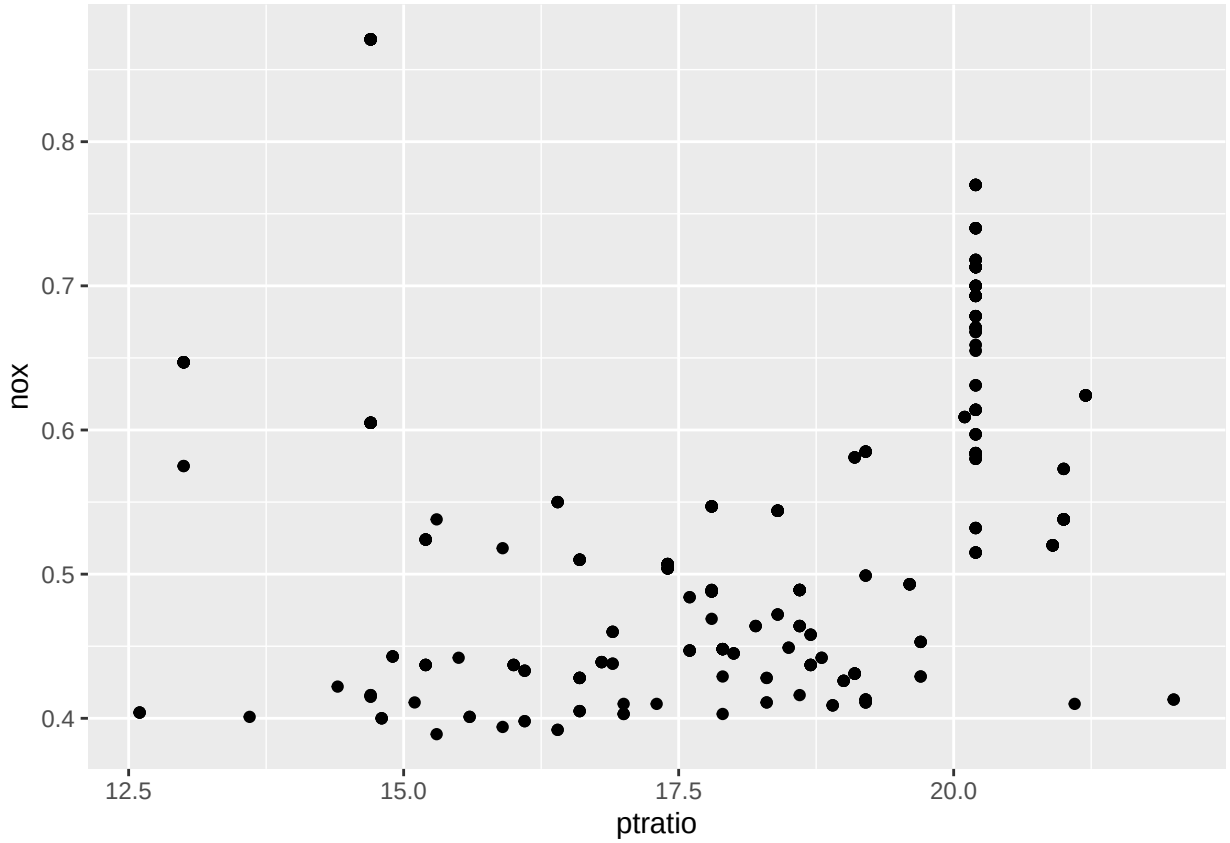
We will drop taxes (because they are 90% correlated with radial highways) and create a new variable, HighPTRatio, so that we can perform a binary analysis.

B. Distributions When we examine histograms we see that a number of variables have distributions that are broken and uneven (zn, indus, nox and rad), suggesting possible hidden groupings. As we will see later, this may lend itself well to decision tree/random forest algorithms. Many of the distributions are also skewed and we can see some likely outliers. However, both models are robust to outliers so we don't do transformations here.



C. Analysis We prepare various datasets for anlaysis:


medv
rm
highPT
0
1
50
40
30
20
10
4
5
6
7
8
9



1. The Full Dataset

```
## * checking for file 'C:\Users\erico\AppData\Local\Temp\RtmpcHWCjX\remotes14d0123974dc\ericonsi-EHData
## * preparing 'EHData':
## * checking DESCRIPTION meta-information ... OK
## * checking for LF line-endings in source and make files and shell scripts
## * checking for empty or unneeded directories
## Omitted 'LazyData' from DESCRIPTION
## * creating default NAMESPACE file
## * building 'EHData_0.1.0.tar.gz'
##
## [1] " "
## [1] "SVM - LINEAR"
## [1] " "
## Support Vector Machines with Linear Kernel
##
## 374 samples
##  11 predictor
##   2 classes: '0', '1'
##
## Pre-processing: centered (11), scaled (11)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 337, 337, 337, 335, 337, 336, ...
## Resampling results across tuning parameters:
##
```

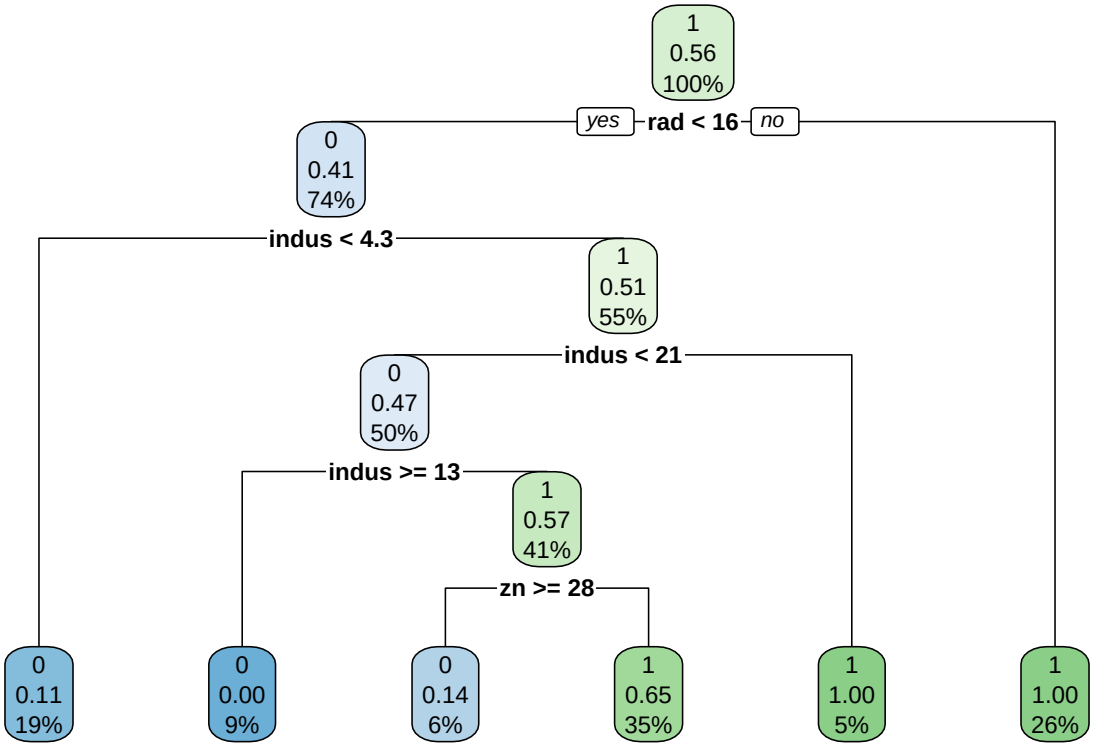
```
##   C          Accuracy   Kappa
##   0.0100000  0.7121038  0.4201827
##   0.1147368  0.7406007  0.4659236
##   0.2194737  0.7360974  0.4580579
##   0.3242105  0.7290076  0.4440101
##   0.4289474  0.7316628  0.4501167
##   0.5336842  0.7352664  0.4568860
##   0.6384211  0.7316865  0.4497515
##   0.7431579  0.7325412  0.4511485
##   0.8478947  0.7307619  0.4470459
##   0.9526316  0.7325400  0.4505215
##   1.0573684  0.7334409  0.4521702
##   1.1621053  0.7352427  0.4559064
##   1.2668421  0.7334646  0.4519200
##   1.3715789  0.7326112  0.4501941
##   1.4763158  0.7317103  0.4482318
##   1.5810526  0.7326112  0.4499971
##   1.6857895  0.7334883  0.4519069
##   1.7905263  0.7352427  0.4554572
##   1.8952632  0.7352202  0.4554292
##   2.0000000  0.7334421  0.4516096
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was C = 0.1147368.
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 29  4
##          1 11 48
##
##                Accuracy : 0.837
##                  95% CI : (0.7454, 0.9058)
##     No Information Rate : 0.5652
##     P-Value [Acc > NIR] : 2.707e-08
##
##                   Kappa : 0.6614
##
##  Mcnemar's Test P-Value : 0.1213
##
##             Sensitivity : 0.7250
##             Specificity : 0.9231
##          Pos Pred Value : 0.8788
##          Neg Pred Value : 0.8136
##              Prevalence : 0.4348
##          Detection Rate : 0.3152
##    Detection Prevalence : 0.3587
##       Balanced Accuracy : 0.8240
##
##        'Positive' Class : 0
##
## [1] " "
## [1] "SVM - RADIAL"
## [1] " "
```

```
## Support Vector Machines with Radial Basis Function Kernel
##
## 374 samples
##  11 predictor
##   2 classes: '0', '1'
##
## Pre-processing: centered (11), scaled (11)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 337, 337, 337, 335, 337, 336, ...
## Resampling results across tuning parameters:
##
##   C     Accuracy   Kappa
##   0.25  0.7560583  0.5057840
##   0.50  0.7738155  0.5391238
##   1.00  0.7970019  0.5855073
##
## Tuning parameter 'sigma' was held constant at a value of 0.1412225
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were sigma = 0.1412225 and C = 1.
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 30  4
##          1 10 48
##
##                Accuracy : 0.8478
##                  95% CI : (0.7579, 0.9142)
##     No Information Rate : 0.5652
##     P-Value [Acc > NIR] : 6.602e-09
##
##                   Kappa : 0.6849
##
##  Mcnemar's Test P-Value : 0.1814
##
##             Sensitivity : 0.7500
##             Specificity : 0.9231
##          Pos Pred Value : 0.8824
##          Neg Pred Value : 0.8276
##              Prevalence : 0.4348
##          Detection Rate : 0.3261
##    Detection Prevalence : 0.3696
##       Balanced Accuracy : 0.8365
##
##        'Positive' Class : 0
##
## [1] " "
## [1] "SVM - POLY"
## [1] " "
## Support Vector Machines with Polynomial Kernel
##
## 374 samples
##  11 predictor
##   2 classes: '0', '1'
```

```
##
## Pre-processing: centered (11), scaled (11)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 337, 337, 337, 335, 337, 336, ...
## Resampling results across tuning parameters:
##
##   degree  scale  C     Accuracy   Kappa
##   1       0.001  0.25  0.5642278  0.00000000
##   1       0.001  0.50  0.5686387  0.01142828
##   1       0.001  1.00  0.6811589  0.31852946
##   1       0.010  0.25  0.7113202  0.41732616
##   1       0.010  0.50  0.7085227  0.41389105
##   1       0.010  1.00  0.7121038  0.42018267
##   1       0.100  0.25  0.7237657  0.43280500
##   1       0.100  0.50  0.7451289  0.47559060
##   1       0.100  1.00  0.7424262  0.47010014
##   2       0.001  0.25  0.5686387  0.01142828
##   2       0.001  0.50  0.6811589  0.31852946
##   2       0.001  1.00  0.7085938  0.40962509
##   2       0.010  0.25  0.7147591  0.42553378
##   2       0.010  0.50  0.7201882  0.43370787
##   2       0.010  1.00  0.7256137  0.43810525
##   2       0.100  0.25  0.7577403  0.50150193
##   2       0.100  0.50  0.7558674  0.50220871
##   2       0.100  1.00  0.7621512  0.51529815
##   3       0.001  0.25  0.6178077  0.14646116
##   3       0.001  0.50  0.6935369  0.36781369
##   3       0.001  1.00  0.7139044  0.42245147
##   3       0.010  0.25  0.7174381  0.42931276
##   3       0.010  0.50  0.7255912  0.43981979
##   3       0.010  1.00  0.7585239  0.50124024
##   3       0.100  0.25  0.7666320  0.52234444
##   3       0.100  0.50  0.7782015  0.54595099
##   3       0.100  1.00  0.7747152  0.53866549
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were degree = 3, scale = 0.1 and C = 0.5.
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 29  3
##          1 11 49
##
##                Accuracy : 0.8478
##                  95% CI : (0.7579, 0.9142)
##     No Information Rate : 0.5652
##     P-Value [Acc > NIR] : 6.602e-09
##
##                   Kappa : 0.6831
##
##  Mcnemar's Test P-Value : 0.06137
##
##             Sensitivity : 0.7250
```

```
##             Specificity : 0.9423
##          Pos Pred Value : 0.9062
##          Neg Pred Value : 0.8167
##              Prevalence : 0.4348
##          Detection Rate : 0.3152
##    Detection Prevalence : 0.3478
##       Balanced Accuracy : 0.8337
##
##        'Positive' Class : 0
##
## [1] " "
## [1] "SVM - RANDOM FOREST"
## [1] " "
## Random Forest
##
## 374 samples
##  11 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 337, 337, 337, 335, 337, 336, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.9390324  0.8784536
##    6    0.9496298  0.8996511
##   11    0.9550352  0.9104231
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 11.
## rf variable importance
##
##            Overall
## indus     100.0000
## rad        95.1269
## nox        47.9100
## dis        42.3349
## medv       25.0556
## zn         14.8095
## rm         12.0677
## age        10.6848
## lstat       9.9190
## crimerate   0.5005
## chas        0.0000
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 39  1
##          1  1 51
##
##                Accuracy : 0.9783
##                  95% CI : (0.9237, 0.9974)
```

```
##     No Information Rate : 0.5652
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9558
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9750
##             Specificity : 0.9808
##          Pos Pred Value : 0.9750
##          Neg Pred Value : 0.9808
##              Prevalence : 0.4348
##          Detection Rate : 0.4239
##    Detection Prevalence : 0.4348
##       Balanced Accuracy : 0.9779
##
##        'Positive' Class : 0
##
## [1] "Parameters:   mtry =  11 , ntree =  500 , nrnodes =  79"
## [1] " "
## [1] "DECISION TREE"
## [1] " "
```




```
## Confusion Matrix and Statistics
##
##           Reference
```

```
## Prediction  0  1
##          0 29  1
##          1 11 51
##
##                Accuracy : 0.8696
##                  95% CI : (0.7832, 0.9307)
##     No Information Rate : 0.5652
##     P-Value [Acc > NIR] : 3.068e-10
##
##                   Kappa : 0.7267
##
##  Mcnemar's Test P-Value : 0.009375
##
##             Sensitivity : 0.7250
##             Specificity : 0.9808
##          Pos Pred Value : 0.9667
##          Neg Pred Value : 0.8226
##              Prevalence : 0.4348
##          Detection Rate : 0.3152
##    Detection Prevalence : 0.3261
##       Balanced Accuracy : 0.8529
##
##        'Positive' Class : 0
##
```

	AUC	Accuracy
Random Forest	0.9778846	0.9782609
Decision Tree	0.8528846	0.8695652
SVM - radial	0.8365385	0.8478261
SVM - poly	0.8336538	0.8478261
SVM - linear	0.8240385	0.8369565

Conclusion

In short, random forest and decision tree algorithms both have their uses. Always reflexively choosing an algorithm because it predicts best is akin to always choosing a Ferrari over a rickety school bus because it goes faster. It's fine until you have to transport 150 crying 6-year-olds to the local zoo. Algorithms don't stand on their own but are used to solve problems, and the nature of the solution needs to match the nature of the problem.